AD_____

Award Number: DAMD17-00-1-0256

TITLE: Molecular Characterization of Resistance

PRINCIPAL INVESTIGATOR: Robert Clarke, Ph.D.

CONTRACTING ORGANIZATION: Georgetown University Medical Center
Washington, DC 20057

REPORT DATE: July 2002

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 2002	Annual (1 Jul 01 - 30 Jun 02)

4. TITLE AND SUBTITLE
Molecular Characterization of Resistance

5. FUNDING NUMBERS
DAMD17-00-1-0256

6. AUTHOR(S):
Robert Clarke, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Georgetown University Medical Center
Washington, DC 20057

E-MAIL:
clarker@georgetown.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

This is an annual report of an IDEA Award to study the molecular mechanisms driving acquired antiestrogen resistance. We have identified several genes associated with resistance to ICI 182,780 (Fulvestrant. Faslodex). Our initial studies have now been published (***Cancer Res***, 62: 3428-3437, 2002). Functional studies on the role of selected genes are in progress. We also have begun building and testing neural network predictors to separate several antiestrogen resistance phenotypes including antiestrogen sensitive, Tamoxifen resistant/ICI 182,780 sensitive, and TAM/ICI 182,780 crossresistant. We also have completed and published a novel method for collecting and processing human biopsies for gene expression microarray studies (***Clin Cancer Res,*** 8: 1155-1166, 2002.). This method allows us to collect samples from patients for future studies of the molecular profiling of antiestrogen responsiveness in patients. Our new algorithm for normalizing gene expression microarray data also has been published in the informatics literature (***IEEE Trans Inf Technol Biomed***, 6: 29-37, 2002).

14. SUBJECT TERMS
breast cancer, estrogens, antiestrogens, drug resistance, hormone resistance, gene expression, microarrays

15. NUMBER OF PAGES
44

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Appendices

1. Wang, Y., Lu, J., Lee, R. & Clarke, R. Iterative normalization of cDNA microarray data. ***IEEE Trans Inf Technol Biomed,*** 6: 29-37, 2002.

2. "Association of Interferon Regulatory Factor-1, Nucleophosmin, Nuclear Factor Kappa- B, and Cyclic AMP Response Element Binding with Acquired Resistance to Faslodex (ICI 182,780)." Gu *et al.,* ***Cancer Res*** 62: 3428-3437, 2002.

3. "Development and Validation of a Method for Using Breast Core Needle Biopsies for Gene Expression Microarray Analyses." Ellis *et al.,* ***Clin Cancer Res*** 8: 1155-1166, 2002.

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

Where copyrighted material is quoted, permission has been obtained to use such material.

Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

PI - Signature Date

Introduction

Antiestrogens have been successfully used in the management of breast cancer since the first clinical trial of Tamoxifen (TAM) in 1971 [1]. TAM produces a significant increase in both overall and recurrence-free survival but resistance almost inevitably arises in most patients [2,3]. We hypothesize that one form of acquired antiestrogen resistance reflects the altered expression of what were previously estrogen-regulated genes. We further hypothesize that only a subset of all estrogen (E2)-regulated genes, those comprising a specific gene network, is responsible for the resistance phenotype. Since TAM (triphenylethylene) and ICI 182,780 (steroidal) induce different ER conformations, we also hypothesize that the consequent patterns of gene regulation will be different and dictate the presence/absence of crossresistance among antiestrogens.

To address these hypotheses, we have generated novel E2-independent and antiestrogen resistant variants of the E2-dependent, MCF-7 human breast cancer cell line (MCF7/MIII, MCF7/LCC1, MCF7/LCC2, MCF-7/LCC9) - recently reviewed in [4]. We also have assembled a panel of additional resistant cells from within this institution and from other investigators. These include additional antiestrogen resistant MCF-7 variants (LY2, R27, R3, MCF-7RR), all of which express ER, and the ER-negative ZR-75-1 (ZR75/LCC3, ZR-75-9a1) and T47D (T47Dco) variants. Other resistance models are currently being obtained from other laboratories or being generated by selection *in vivo* selection against TAM in athymic nude rats (rats and humans perceive TAM as a partial agonist, mice perceive TAM as a pure agonist).

This is an Idea Award to study the genes and patterns of genes expressed in acquired antiestrogen resistance in cell culture models. The PI will apply new, state-of-the-art technologies to identify key endocrine-regulated molecular pathways to apoptosis/proliferation. By identifying key components of these pathways, we may be able to predict response to first-line and crossover antiestrogenic therapies, and/or provide novel therapeutic strategies for antiestrogen resistant tumors.

Body of Text

Our purpose is to evaluate a series of antiestrogen responsive and resistant breast cancer cell lines for their patterns of gene expression. We will explore these data, using state-of-the-art clustering pattern analysis through joint use of the standard Finite Normal Mixture models and probabilistic component subspaces, where the multimodal clusters will be automatically identified using Akaike information criterion and Minimal Description Length analyses. We also will apply the more computationally simplistic methods used by others in the field.

In our previous report, we made one change to the specific aims and Statement of Work. Our collaborations with Dr. Wang's group at Catholic University of America have increased substantially, and we have begun to develop and test several new algorithms for mining the high dimensional data sets produced by gene expression microarray analyses.

Specific Aims

Specific Aim 1: use gene microarrays to identify differentially expressed genes in a panel of breast cancer cell lines.

Specific Aim 2: explore the data from Aim 1 to identify those differentially expressed gene clusters most closely associated with acquired antiestrogen resistance and test further novel algorithms for the analysis of gene expression microarray data.

Specific Aim 3: begin to assess the likely functional relevance of representative members of these clusters and study their expression in human breast cancer biopsies.

Long term aims: establish a pattern(s) of gene clusters that can predict antiestrogen responses in patients. This could lead to a more effective identification of candidates for specific antiestrogen therapies and identify those patients least likely to respond and who may benefit from an early initiation of cytotoxic chemotherapy.

Statement of Work and Progress on the Work Proposed

The Specific Aims of this application are being addressed in the studies outlined in the Statement of Work.

TASK 1: *Use gene microarrays to identify differentially expressed genes in a panel of breast cancer cell lines.*

A. Expand cells and prepare RNA from cell lines for pilot study
B. Label RNA populations, probe microarrays and digitize data
C. Optimize probing/reprobing as necessary
D. Expand cells and prepare RNA from replicate cultures of remaining cell lines (including ER-negative cells) for the baseline study
E. Label RNA populations, probe microarrays. and digitize data
F. Expand cells, treat with ICI 182,780 and 4-hydroxyTAM and prepare RNA from replicate cultures
E. Label RNA populations, probe microarrays, and digitize data

We have effectively completed this aim and have generated all the treated cell populations and RNAs. We have arrayed most of the populations on the initial Research Genetics arrays and individually aligned all of the digitized images. We used Pathways vs. 4.0 and independently align each of the ~4,000 spots/array; this is rather time consuming but provides much higher quality data than using only the software to align automatically each spot. In year 3, we will

perform limited studies on the same RNAs using our Affymetrix system and compare data across platforms.

We have one paper on our initial studies from the Clontech arrays (Gu et al., *Cancer Res* 62: 3428-3437, 2002). A paper on the data from the Research Genetics platform is in preparation.

TASK 2: *Explore the data from Aim 1 to identify those differentially expressed gene clusters most closely associated with acquired antiestrogen resistance.*

A. Perform preliminary analysis of pilot study and identify candidates for further study
B. Generate reagents and confirm differential regulation/expression of candidates from the pilot study
C. Analyze the data from the baseline study (includes evaluation of ER-negative models both separately and together with ER-positive cell) using all four data analysis approaches and identify candidates for further study
D. Generate reagents and confirm differential regulation/expression of candidates from the baseline study
E. Analyze the data from the treatment study using all four approaches and identify candidates for further study
F. Perform overall and final analyses, compare data from each analytical method and identify candidates for further study
G. Generate reagents and confirm differential regulation/expression of candidates from the treatment study
H. Test novel algorithms for the analysis of gene expression microarray data

We have completed "A" and "B" and (in Aim 1) and generated most of the data/reagents needed for "C" and "D".

Our initial studies identified several genes of interest. The paper by Gu *et al.* describes several of these for which we have already confirmed differential expression and/or antiestrogenic regulation. Other genes in this paper will be evaluated under "D". Our data in this paper includes both microarray (Table 1) and serial analysis of gene expression (SAGE; Table 2) data. Since our study is hypothesis driven rather than technology driven, we will include genes identified by SAGE in our future studies.

Table 1 *Representative list of differentially expressed genes identified by gene microarray analyses*

Gene[a]	Unigene #	MCF7/LCC1[b]	MCF7/LCC9	Gene Function
NFκB	Hs.75569	1	2	transcription factor involved in cell survival signaling
SOD	Hs.75428	1	2	enzyme involved in detoxifying oxygen radicals
EGR-1	Hs.326035	3	1	transcription factor
EGFR	Hs.77432	2	1	growth factor receptor
IRF-1	Hs.80645	2	1	transcription factor involved in signaling to cell cycle arrest and apoptosis
TNFα	Hs.241570	2	1	Cytokine
TNF-R1	Hs.159	2	1	cytokine receptor involved in signaling to apoptosis

[a]Abbreviations are NFκB, nuclear factor kappa B; SOD, superoxide dismutase; EGR-1, early growth response gene-1; EGF-R, epidermal growth factor receptor; IRF-1, interferon regulatory factor -1; TNFα, tumor necrosis factor alpha; TNF-R1,tumor necrosis factor-receptor 1.

[b]Data are represented as level of expression relative to the other cell line. Data are based on the mean values for each gene (6 microarrays of MCF7/LCC1; 5 microarrays of MCF7/LCC9). Values are expressed to the nearest integer.

Table 2 *Differentially expressed genes identified in the MCF7/LCC1 and MCF7/LCC9 SAGE libraries*

Putative Gene[a]	Unigene #	MCF7/ LCC1	MCF7/ LCC9	Differ-ence[b]	p-value[c]	Gene Function
N-ras related gene	Hs.260523	2	20	10-fold	<0.001	G-protein
Cathepsin D	Hs.343475	7	34	5-fold	<0.001	protease involved in tumor invasion
X-box binding protein-1	Hs.149923	7	25	4-fold	<0.001	transcription factor
Prefoldin 5	Hs.288856	6	21	4-fold	0.002	chaperone for unfolded proteins
HSP-27	Hs.76067	23	55	2-fold	0.001	stress response protein
Vit B-12 binding protein	Hs.2012	17	37	2-fold	0.002	vitamin binding protein
Nucleophosmin[d]	Hs.9614	10	14	1.5-fold	>0.05	oncogenic nucleolar protein
L14	Hs.738	13	2	6-fold	0.021	ribosomal protein
Death associated protein-6	Hs.336916	11	2	6-fold	0.049	apoptosis associated protein
EF-γ	Hs.2186	22	6	4-fold	0.014	translation elongation factor
Ferritin, heavy polypeptide-1	Hs.62954	54	16	3-fold	<0.001	iron binding protein

[a] The gene designations are considered putative, although, the identity of most genes designated in this fashion have been shown to be correct. These genes include those Tags where: (a) the fold difference is ≥2-fold, (b) the Tag could represent ≤2 genes, and (3) represents ≥0.10% of either the MCF7/LCC1 and/or MCF7/LCC9 SAGE library.

[b]Predicted fold difference in gene expression between MCF7/LCC1 *vs.* MCF7/LCC9 cells.

[c]Obtained by χ^2 analyses; p-values estimated to 3 significant figures.

[d] NPM (not statistically significant) is shown because we know it to be both estrogen regulated and associated with TAM treatment in patients.

We have already trained and tested our initial neural predictors of antiestrogen resistance, using the data from the Research Genetics platform. Since these studies are not finished, we will present the mature data in our next annual report. We hope also to have published this predictor before the next report and to include a preprint in the report. The genes in this predictor, or other differentially expressed genes associated with the phenotypes, will be the candidate genes for "F" and "G".

In our last report we presented a new normalization algorithm; this study has now been published (*IEEE Trans Inf Technol Biomed,* 6: 29-37, 2002). We also have developed a novel "block principal components analysis" method for exploring gene expression microarray data. A manuscript has been submitted and the merthod and a reprint will be included in the next report. Thus, we continue to address successfully Task 2H.

TASK 3: *Begin to assess the likely functional relevance of representative members of these clusters and study their expression in human breast cancer biopsies.*

A. Obtain/generate reagents for the 1-2 candidates from the pilot study
B. Initiate pilot studies using transient transfection analyses
C. Initiate functional (transient) studies of candidates from baseline study
D. Initiate functional (stable transfection) studies of candidates from baseline study
E. Initiate functional (transient) studies of candidates from treatment study
F. Initiate functional (stable transfection) studies of candidates from overall analysis (only if new candidates are identified)

We continue to investigate the functional relevance of those genes/proteins that receive sufficient priority. We cannot perform detailed functional studies of all our candidates within this application but have used the present DOD award to obtain preliminary data to support requests for funding to study specific genes. In this regard, we successfully used the preliminary data generated on XBP-1 to attract additional DOD funding to perform detailed mechanistic and translational studies of XBP1. Thus, we are now able to study XBP-1 in the broader context of its role in endocrine signaling in breast cancer, but with a focus on its potential contribution to acquired estrogen-independence and antiestrogen resistance. This successful application also includes retrospective studies to identify the prognostic role of XBP-1 in breast cancer (outcome independent of therapy) and to assess whether it may have predictive relevance in improving the ability to predict which patients are most likely to respond to endocrine therapies.

We have completed studies showing that antiestrogen resistant MCF7/LCC9 cells, which overexpress NFκB transactivation (promoter-reporter activity), are more sensitive to the growth inhibitory effects of Parthenolide, a specific inhibitor of NFκB. Growth inhibition was assessed using a dye-based assay that effectively estimates cell number. These data are consistent with our hypothesis that increased NFκB activation in these cells contributes to their ability to survive prolonged antiestrogen exposure. These data are published in Gu et al., *Cancer Res* 62: 3428-3437, 2002).

To maintain focus within this application, we have limited our initial studies to NFkB and IRF-1. Our intention is to obtain sufficient preliminary data to support an R01 or DOD application focused on

these two genes and their interactions in antiestrogen resistance. We have continued to study the role of our dominant negative interferon regulatory factor-1 (dnIRF-1). Our studies progressed somewhat more slowly than expected, mostly due to the Fellow performing the studies taking maternity leave. However, we hope to be back on track within the next few months and complete the few remaining experiments required to solidify and extend the data presented in last year's report. We will then submit a manuscript on dnIRF-1 (within the next 12 months). The mature data will be included in our next annual report.

Key Research Accomplishments (bulleted)

- Completed and published manuscript describing data from gene microarray and SAGE studies based on the data presented in the previous report. These data show the altered regulation of X-box binding protein-1, NFκB, NPM and IRF-1 in acquired antiestrogen resistance (manuscript submitted).
- Completed collection of RNA from resistant and parental cell cultures.
- Obtained microarray data from resistant and parental cell cultures.
- Completed microarray data preprocessing and confirmed data quality.
- Built initial neural predictors - will be completed within the next few months.
- Completed studies implicating NFkB as a mediator of survival from prolonged antiestrogen exposure.
- Completed and published a new algorithm based on regression through the origin for normalizing gene expression microarray data.
- Completed and published a pilot study showing our ability to generate accurate predictive neural networks based on gene expression microarray data. The neural network predictors that can accurately identify the phenotype of unknown samples as being cancer or noncancer.

Reportable Outcomes

Reportable outcomes are presented as manuscripts and abstracts.

Manuscripts and Abstracts

We have published several studies directly related to the funded work.

Manuscripts

1. Ellis, M., Davis, N., Coop, A., Liu, M., Schumaker, L., Lee, R.Y., Srikanchana, R., Russell, C., Singh, B., Miller, W.R., Stearns, V., Pennanen, M., Tsangaris, T., Gallagher, A., Liu, A., Zwart, A., Hayes, D.F., Lippman, M.E., Wang, Y. **& Clarke, R.** "Development and validation of a method for using breast core needle biopsies for gene expression microarray analyses." *Clin Cancer Res,* 8: 1155-1166, 2002.

2. Wang, Y., Lu, J., Lee, R. **& Clarke, R.** "Iterative normalization of cDNA microarray data ." *IEEE Trans Inf Techol Biomed,* 6: 29-37, 2002.

3. Gu, Z., Lee, R.Y., Skaar, T.C., Bouker, K.B., Welch, J.N., Lu, J., Liu, A., Davis, N., Leonessa, F., Brünner, N., Wang, Y. **& Clarke, R.** "Association of interferon regulatory factor-1, nucleophosmin, nuclear factor-κB and cAMP response element binding with acquired resistance to Faslodex (ICI 182,780)." *Cancer Res,* 8: 1155_1166, 2002.

4. Welch, J.N. **& Clarke, R.** "ErbB-2 expression and drug resistance in cancer." *Signal*, in press (review).

Reprints of papers #1-3 are included in the appendix.

Abstracts

1. Welch, J.N., Chrysogelos, S. **& Clarke, R.** "Expression and function of the epidermal growth factor receptor in breast cancer cells exposed to chemotherapy." *Proc Am Assoc Cancer Res* 42: 938, 2001.

2. Bouker, K.B., Skaar, T.C., Fernandez, D. **& Clarke, R.** "Antiestrogens regulate IRF-1 expression in sensitive but not resistant breast cancer cells." *Proc Am Assoc Cancer Res* 43: 761, 2002.

3. Zhu, Y., Bouker, K., Skaar, T., Zwart, A., Gomez, B., Hewitt, S., Singh, B., Liu, A. **& Clarke, R.** "High throughout tissue microarray assessment of expressions of progression-related genes - NFkB, nucleophosmin, X-box binding protein-1 and IRF-1 in breast cancer." *Proc Am Assoc Cancer Res* 43: 762, 2002.

We also presented our data at the recent DOD meeting in Orlando, FL.

Conclusions

We have made good progress in our studies on the molecular characterization of antiestrogen resistance is evident in our productivity as measured by publications and new methods and preliminary data. The study is on-track and the amount of data accumulating is considerable. However, several new algorithms underdevelopment are showing good performance in our very preliminary analyses of published high dimensional data sets. Our data with NFκB, IRF-1 and the dnIRF-1 are encouraging

and suggest we may be on the right track to identifying new signal transduction pathways associated with acquired antiestrogen resistance. For example, these data show that resistant cells are more sensitive to inhibition of NFκB. Overexpression of IRF-1, which is suppressed by estrogens and induced by antiestrogens, is associated with reduced cell proliferation. The dnIRF-1 provide an opportunity to further explore some of the mechanistic effects of this gene in acquired antiestrogen resistance. We also have been successful in using the preliminary data generated in this application to attract other funding.

Literature Cited

1. Cole, M. P., Jones, C. T. A., and Todd, I. D. H. A new antioestrogenic agent in late breast cancer. An early clinical appraisal of ICI 46474. Br J Cancer, *25*: 270-275, 1971.
2. EBCTCG Early Breast Cancer Trialists Collaborative Group: Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. Lancet, *399*: 1-15, 1992.
3. EBCTCG Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomized trials. Lancet, *351*: 1451-1467, 1998.
4. Clarke, R., Leonessa, F., Welch, J. N., and Skaar, T. C. Cellular and molecular pharmacology of antiestrogen action and resistance. Pharmacol Rev, *53*: 25-71, 2001.

Iterative Normalization of cDNA Microarray Data

Yue Wang, Jianping Lu, Richard Lee, Zhiping Gu, and Robert Clarke

Abstract—**This paper describes a new approach to normalizing microarray expression data. The novel feature is to unify the tasks of estimating normalization coefficients and identifying control gene set. Unification is realized by constructing a window function over the scatter plot defining the subset of constantly expressed genes and by affecting optimization using an iterative procedure. The structure of window function gates contributions to the control gene set used to estimate normalization coefficients. This window measures the consistency of the matched neighborhoods in the scatter plot and provides a means of rejecting control gene outliers. The recovery of normalizational regression and control gene selection are interleaved and are realized by applying coupled operations to the mean square error function. In this way, the two processes bootstrap one another. We evaluate the technique on real microarray data from breast cancer cell lines and complement the experiment with a data cluster visualization study.**

Index Terms—**Data normalization, dynamic programming, gene expression, gene microarray, linear regression.**

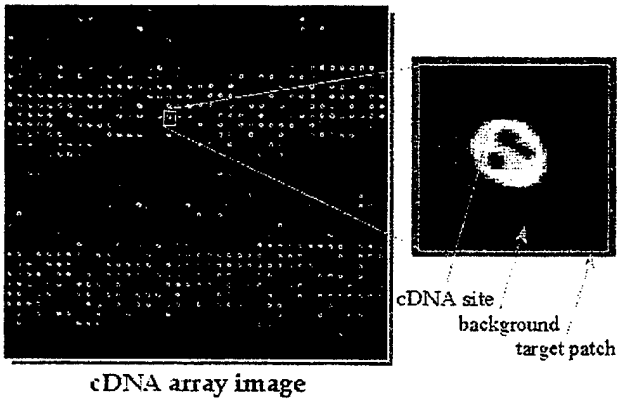


Fig. 1. Example of cDNA microarray image.

I. Introduction

SPOTTED cDNA microarrays are emerging as a powerful and cost-effective tool for the large-scale analysis of gene expression. Using this technology, the relative expression levels in two or more mRNA populations derived from tissue samples can be assayed for thousands of genes simultaneously [1], [2]. Microarrays are potentially powerful tools for investigating the mechanism of drug action. Two recent studies have described the application of high-density microarrays to examine the effects of drugs on gene expression in yeast as a model system. A similar method applied to human breast cancer cells and tissues would have direct utility in the identification and validation of novel therapeutics. It is widely accepted that the pattern of genes expressed within a specific cell is essentially responsible for its phenotype. The most widely publicized use of gene microarrays has been in cancer research.

From a statistical point of view, sources of measurement error within an array, and variation between arrays, must be quantified and taken onto account in order to make indirect comparisons among samples that have not been directly assayed on the same array. For example, gene microarrays vary with production batches, e.g., introducing variations in the amount of probe that hybridizes to areas of the support that do not contain target cDNAs, or the amount of the cDNA spotted onto the support surface. The specific activity of the probe will vary from probe to probe, often reflecting variations in the amount of signal produced by each molecule of label incorporated into the probe.

Two major data preprocessing operations are involved: background correction and interexperiment normalization. In background correction, local sampling of background can be used to specify a threshold that a true signal must exceed. It is even possible to accurately detect weak signals and extract a mean intensity above background for the target [3]. A typical cDNA array image is given in Fig. 1.

In carrying out comparisons of expression data using measurements from a single array or multiple arrays, the question of normalizing data arises. A reasonable assumption, adopted by most researchers, is that all experiments are carried out under conditions of a large excess of immobilized probe relative to labeled target. The kinetics of hybridization are therefore pseudofirst order, and interprobe competition is not a factor [3]. Under these assumptions, the linear differences arising from the exact amount of applied target, extent of target labeling, efficiencies of fluor excitation and emission, and detector efficiency can be compounded into a single variable. Two major strategies can be used to carry out normalization. One is based on a consideration of all of the genes in the sample, and the other on a designated subset expected to be unchanged over most circumstances, called the control gene set. In instances of closely related samples, global normalization (e.g., using all genes) will be a useful tool. As samples become more divergent, a good normalization may be achieved using a subset of constantly expressed genes (e.g., using only control genes) [3].

The work most closely related to our methodology was reported in [4]. The authors introduced a comparison of gene expression levels arising from cohybridized samples by taking ratios of average expression levels for individual genes. A novel method of image segmentation was presented to identify cDNA

Manuscript received March 28, 2001; revised September 14, 2001. This work was supported in part by the National Institutes of Health under Grants 5R21CA83231 and R01CA/AG58022.

Y. Wang and J. Lu are with the Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064 USA.

R. Lee and R. Clarke are with the Lombardi Cancer Center, Georgetown University Medical Center, Washington, DC 20007 USA.

Z. Gu is with the Celera Genomics, Inc., Rockville, MD 20850 USA.

Publisher Item Identifier S 1089-7771(02)02009-5.

target sites, and a hypothesis test and confidence interval was developed to quantify the significance of observed differences in expression ratios. In particular, the probability density of the ratio and the maximum-likelihood estimator for the distribution were derived, and an iterative procedure for signal calibration was developed. In general, however, an integral of ratios is not the same as a ratio of integrals, and simple ratios of the data will not necessarily provide unbiased estimates of expression ratios. Alternatively, the mean value of all signals on the hybridized filter can be used for normalization, and further normalizations can be done to a reference hybridization [5]. Nonetheless, the optimal approach remains controversial.

II. Method and Algorithm

In this paper, we adopt a somewhat different approach to the problem of normalizing microarray expression data. Rather than rejecting those control genes that give rise to a large normalization error, we attempt to iteratively correct them. In a nutshell, our idea is to bootstrap by alternating between estimating normalization coefficients and identifying control gene subset. The framework is furnished by constructing a window function over the scatter plot defining the subset of constantly expressed genes. Specifically, this window measures the consistency of the matched neighborhoods in the scatter plot and provides a means of rejecting control gene outliers. We evaluate the technique on real microarray data from breast cancer cell lines and complement the experiment with a data cluster visualization study.

Our goal is to generate a transformation that best maps the expression levels of floating data set onto their counterparts in a reference data set. Assume that data points $\{x_1, x_2, \ldots, x_{n_c}\}$ and $\{y_1, y_2, \ldots, y_{n_c}\}$ are the expression levels of the control or housekeeping genes from two microarray experiments, where n_c is the total number of control genes. In this paper, we use $\{x_i\}$ as the floating data set and $\{y_i\}$ as the reference data set. We further assume that the normalization can be accurately achieved through a linear regression mapping

$$y_i = ax_i + b \tag{1}$$

where a is the true ratio of the data and b is the bias correction of the data. Since there are two free parameters in the transformation, the estimation of their values requires a minimum of two data points that are known to be in correspondence. By considering noise effect, however, more control points are needed to produce an accurate estimate. This process is overconstrained and can be solved using least squares estimation. Clearly, a natural criterion is the minimum mean squared error between the two control data subsets. Based on the expression levels of the control genes, the mean squared error (MSE) can be written as

$$\varepsilon = \frac{1}{n_c}\sum_{i=1}^{n_c}[y_i - (ax_i + b)]^2. \tag{2}$$

Thus, the search principle for estimating the optimal values of a and b is simply taking the partial derivatives of the MSE and setting them to zero. It can be shown that the estimated linear regression coefficients a and b can be calculated by [6]

$$a = \frac{\sum_{i=1}^{n_c}(x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^{n_c}(x_i - \mu_x)^2} \tag{3}$$

$$b = \mu_y - a\mu_x \tag{4}$$

where μ_x and μ_y are the means of $\{x_i\}$ and $\{y_i\}$, respectively, given by

$$\mu_x = \frac{1}{n_c}\sum_{i=1}^{n_c} x_i, \qquad \mu_y = \frac{1}{n_c}\sum_{i=1}^{n_c} y_i \tag{5}$$

and the normalization shall be performed for all of the data points in the floating data set based on (1).

The accuracy of the method highly depends on the selection of control genes. In addition to the predetermined control genes, including housekeeping genes, we shall add more control genes based on a reasonable heuristics that the genes that are nondifferentially expressed should be considered as control genes in normalization. Posed in this way, there is a basic "chicken-and-egg" problem [7]. Before a good control gene subset can be defined, expression levels of all genes need to be reasonably normalized. Yet, this normalization is, after all, the ultimate goal of computation.

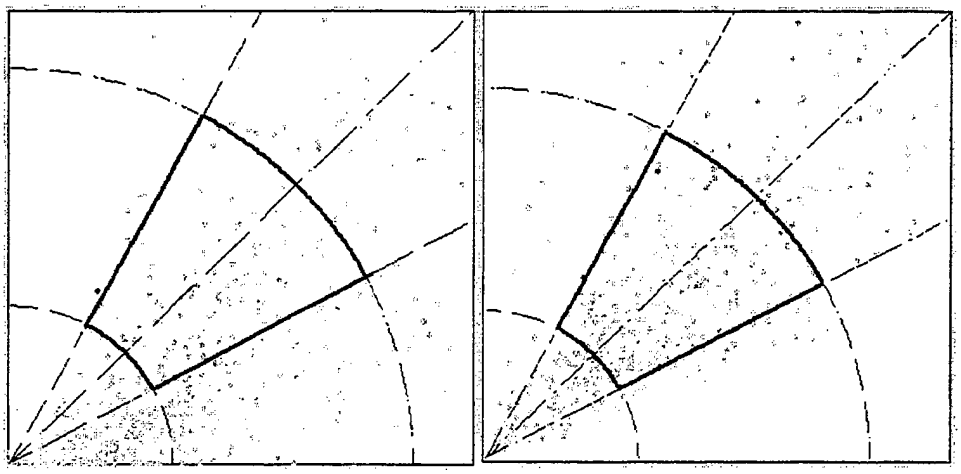
We propose an iterative regression normalization algorithm to solve this problem. First, solely based on the predetermined control genes such as housekeeping genes, we will conduct an initial normalization to all data sets based on (1)–(5). Since an accurate data analysis requires several repetitive cDNA hybridizations in microarray studies [8], starting from the whole data set, we will then eliminate those genes from the control gene list whose expressions have a large standard deviation across replications, namely, outliers, according to the criterion given by

$$\mu_i = \frac{1}{m_i}\sum_{j=1}^{m_i} x_{i,j}, \qquad \sqrt{\frac{1}{m_i}\sum_{j=1}^{m_i}\left(\frac{x_{i,j}}{\mu_i} - 1\right)^2} \geq \epsilon_1 \tag{6}$$

for all genes, where m_i is the number of replications for gene i in the experiment, $x_{i,j}$ is the expression level of gene i in the jth replication, μ_i is the mean of replications, and ϵ_1 is a predetermined threshold.

In our experiment, ϵ_1 is determined as follows. For each of the genes, the replications are normalized by its mean and the normalized standard deviation is calculated. A mean standard deviation is then obtained by the sample average of the individual normalized standard deviations. Our experience has shown that ϵ_1 being two times of the mean standard deviation is appropriate and effective. It should be noted that this criterion will also eliminate differentially expressed genes from the control gene list. Thus, a gene will be selected as a control gene if its expression level pair across reference and floating experiments satisfies

$$\epsilon_2 \leq \sqrt{x_i^2 + y_i^2} \leq \epsilon_3 \quad \text{and} \quad \left|\frac{\log x_i}{\log y_i} - 1\right| \leq \epsilon_4 \tag{7}$$



(a) (b)

Fig. 2. Example of control gene selection window. (a) Before and (b) after normalization.

where ϵ_2, ϵ_3, and ϵ_4 are the empirically predetermined thresholds defining the subset of constantly expressed genes. It can be seen that (7) defines a window function over the scatter plot. A typical window function is illustrated in Fig. 2. In particular, as we have noted according to our experience, ratios can be very unstable when one (or both) of the signals is small or large. Thus, we further eliminate unstably expressed genes from the control gene list using the constraints defined by ϵ_2 and ϵ_3. Clearly, ϵ_4 provides the boundaries of a constantly expressed gene subset.

The algorithm first generates the interim scatter plot of the data sets through the observations and the current parameter estimates [(1)] and then updates parameter estimates using a newly defined control gene subset [(3)–(5)]. The procedure cycles back and forth between these two steps until it reaches a stationary point where no significant change occurs to the content of the control gene subset. A summary of the major steps is given as follows.

1) Based on predetermined control genes including housekeeping genes, estimate initial values of $a^{(0)}$ and $b^{(0)}$ and perform an initial normalization using (3)–(5) and (1), where only one data set is used as a reference set and all other data sets are considered as floating sets and shall be normalized to the reference set.
2) Eliminate those genes from the control gene list whose expressions have a large standard deviation across replications, according to the criterion given by (6).
3) For each of experiment pairs, construct a new control gene subset by selecting additional control genes that satisfy (7).
4) Based on the newly constructed control gene subset, estimate interim values of $a^{(m)}$ and $b^{(m)}$ and perform data normalization for each of the floating data sets using (3)–(5) and (1), where m is the iteration index.
5) Repeat Steps 3) and 4) until the convergence ($a^{(\infty)} \rightarrow 1$ and $b^{(\infty)} \rightarrow 0$) is reached or no significant change occurs to the content of the control gene subset.

The philosophy for estimating normalization coefficients and identifying a control gene set is similar in spirit to the *self-organization* principle [9], [10]. The structure of window function gates contributions to the control gene subset used to estimate normalization coefficients such that possible oscillation during algorithm convergence can be prevented. Specifically, the window function defines a neighborhood of scatter centroid to gating consistency contribution of the control gene subset to normalization. By making the value ϵ_4 of the topological window function decrease with time, the neighborhood is initially very large and shrinks slowly to its final desired size (e.g., a nearest neighbor structure). A popular choice for the dependence of ϵ_4 on discrete time m is the exponential decay [9]. In addition, the actual algorithm implementation concerns the issue of numerical stability. We have applied a simple dynamic programming technique to estimating normalization coefficients, called a factoring-shifting (FS) procedure.

F-Step

$$a^{(k|m)} = \frac{\sum_{i=1}^{n_c} x_i^{(k|m)} y_i}{\sum_{i=1}^{n_c} \left[x_i^{(k|m)}\right]^2} \tag{8}$$

$$x_i^{(k'|m)} = a^{(k|m)} x_i^{(k|m)} \tag{9}$$

S-Step

$$b^{(k|m)} = \frac{1}{n_c} \sum_{i=1}^{n_c} \left[y_i - x_i^{(k'|m)}\right] \tag{10}$$

$$x_i^{(k+1|m)} = x_i^{(k'|m)} + b^{(k|m)} \tag{11}$$

where at each complete cycle of the procedure, we first use the "old" set of floating data to determine the normalization factor $a^{(k|m)}$ using (8) by setting $b = 0$ and simply rescale floating data values using (9). These interim results $x_i^{(k'|m)}$ are then used to obtain the normalization shift $b^{(k|m)}$ using (10) by setting

$a = 1$ and further generate "new" values $x_i^{(k+1|m)}$ of floating data using (11). The procedure cycles back and forth until the values of $a^{(\infty|m)}$ and $b^{(\infty|m)}$ reach their stationary points.

In relation to previous work, the concept of using linear regression analysis for microarray normalization can be traced back to [4] and was further developed in [12] for iterative regression in conjunction with control gene selection. Such approaches are based on several assumptions regarding the data and can be considered as special cases of our framework [5].

The primary assumption is that for either the entire collection of arrayed genes or some subset such as housekeeping genes, the shift of the measured expression averaged over the set is zero (e.g., $b = 0$) and the ratio of normalized expression pair averaged over the set should be one [e.g., $1/n_c \sum_{i=1}^{n_c}(y_i/ax_i) = 1$]. Under these assumptions, there are basically three major approaches for calculating the normalization factor a [5]. The first simply uses the mean value of all the background-corrected signals. Normalization can be separately performed for each of the data sets, without explicitly calculating a and selecting a reference data set [15]. Specifically, if a raw data pair is denoted by $(\{x_i\}, \{y_i\})$, normalization leads to $(\{x_i/\sum_{i=1}^{n_c} x_i\}, \{y_i/\sum_{i=1}^{n_c} y_i\})$. By multiplying the pair with $\sum_{i=1}^{n_c} y_i$, the result is equivalent to using (4), that is, $(\{ax_i\}, \{y_i\})$, where $a = \sum_{i=1}^{n_c} y_i/\sum_{i=1}^{n_c} x_i$ and $b = 0$. A second approach uses simplified linear regression analysis, called linear regression through the origin [6]. Consequently, a scatter plot of the normalized data set pair should have a slope of one [5]. By setting $b = 0$ in (1) and (2), the normalization factor is given by $a = \sum_{i=1}^{n_c} x_i y_i/\sum_{i=1}^{n_c} x_i^2$. A third approach relies on the assumption that, for control genes, the distribution of expression levels can be modeled and the mean of the ratio adjusted to one [4]. An iterative procedure was developed to estimate a by $1/n_c \sum_{i=1}^{n_c}(y_i/x_i)$, once again setting $b = 0$. It should be noticed that some heuristic approximations have been made in using these approaches, since, in general

$$\frac{\sum_{i=1}^{n_c} y_i}{\sum_{i=1}^{n_c} x_i} \neq \frac{\sum_{i=1}^{n_c} x_i y_i}{\sum_{i=1}^{n_c} x_i^2} \neq \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{y_i}{x_i} \tag{12}$$

and

$$b \neq 0 \tag{13}$$

while our method is presented as a standard linear regression analysis without any approximation step.

III. Experiment and Discussion

In this section, we will provide experimental evaluation of our new normalization method. This investigation has two related strands. First, we will furnish examples demonstrating the use of an iterative normalization scheme on real microarray data. Here, we will use two different data sets. The first of these involves *within-class* normalization of data from LCC1 breast cancer cell lines across replications. The second example involves normalizing *between-class* breast cancer cell line data from LCC1 against LCC9, whose phenotypes are known to be different from LCC1.

In the second strand of our experiments, we will provide an algorithm accuracy analysis. Here, we confine ourselves to the linear regression variant of the normalization process. The aim is to experimentally compare our iterative algorithm with the performance of each of its components taken individually, thus to demonstrate that the combined processing of both control gene selection and transformation coefficient estimation yields significant advantages over existing methods. In addition, we would like to acknowledge that although the cell lines are not fully representative of solid tumors in humans, their patterns of gene expression profile are rich in information with respect to drug resistance.

We obtained gene expression profiles from two breast cancer cell lines. MCF7/LCC1 is an estrogen-independent but antiestrogen responsive variant of the MCF-7 human breast cancer cell line [14], [15]. An antiestrogen resistant variant (MCF7/LCC9) was obtained by stepwise selection of MCF7/LCC1 cells against the steroidal antiestrogen ICI 182 780 (trade name: Faslodex). MCF7/LCC9 cells have many of the characteristics seen in antiestrogen-resistant human breast cancers and provide a novel model in which to study antiestrogen resistance [14].

Gene expression profiles were obtained using the AtlasTM Human Array cDNA expression microarrays (Clontech, Laboratories, Inc., Palo Alto, CA). These microarrays are produced on nylon filters and contain 588 target genes and nine housekeeping genes. Briefly, total RNA was obtained from independent cultures of MCF7/LCC1 and MCF7/LCC9 cells with the TRIzol reagent (Life Technologies, Grand Island, NY). One μg of DNase-treated mRNA was primed with Clontech's cDNA Synthesis Primer mix and the product reverse transcribed into radiolabeled cDNA with [−32P] dATP (Amersham Life Science Inc., Arlington Heights, IL). Probes were purified, denatured, and both C0t-1 DNA and 1 M NaH2PO4 (pH 7.0) added to the denatured probe. Each microarray was prehybridized with 5-ml ExpressHyb buffer and 0.5-mg denatured DNA from sheared salmon testes. Microarray filters were hybridized overnight with the appropriate [−32P]-labeled cDNA probe. The array was extensively washed and sealed in plastic, with signals detected by phosphorimage analysis using a Molecular Dynamics Storm phosphorimager (Molecular Dynamics, Sunnyvale, CA). Digitization of these signals provided numerical values representing the signal for each gene.

Generally, it has been assumed that, under variable conditions, the expression of housekeeping genes remains unchanged. Hence, high-throughput differential expression data can rely on these genes for data normalization. However, recent data indicate deviation from this concept [11].

To assess the effectiveness of housekeeping genes in normalizing cDNA microarray data, a normalization based on single linear regression is performed using only the set of nine housekeeping genes suggested by CLONTECH. The scatter plots of normalization results are given in Fig. 3. Although log–log-based scatter plots are widely used, we have decided to use original scaled scatter plots since our numerical simulations have shown possible misleading perceptions from the "distorted" shape of actual data distribution. Particularly focusing on breast cancer, we have observed significant variations in

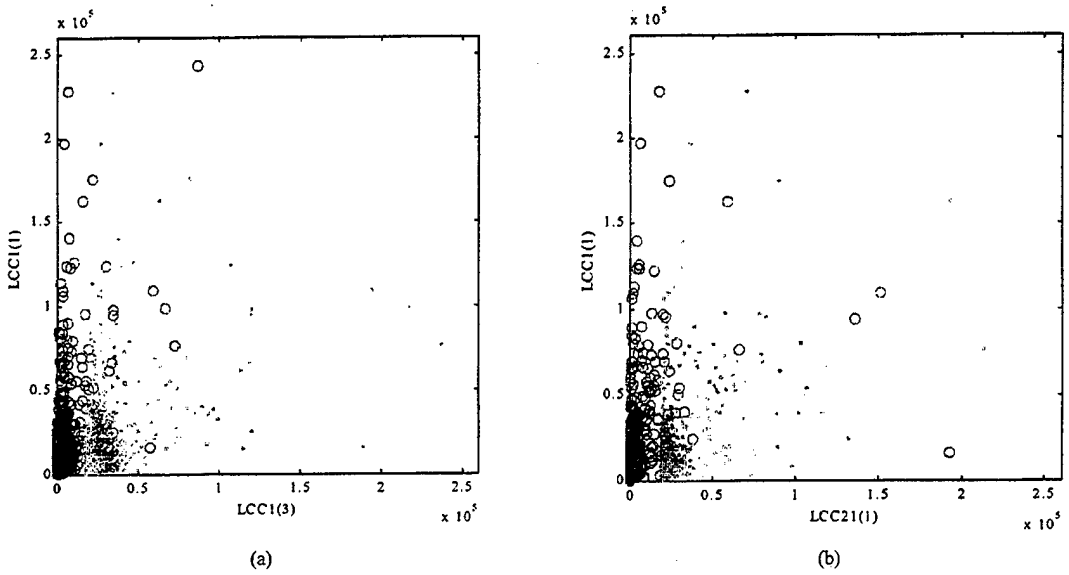


Fig. 3. (a) Scatter plot of *within-class* normalized microarray data based on nine housekeeping genes. (b) Scatter plot of *between-class* normalized microarray data based on nine housekeeping genes. (Circle: before normalization; dot: after normalization.)

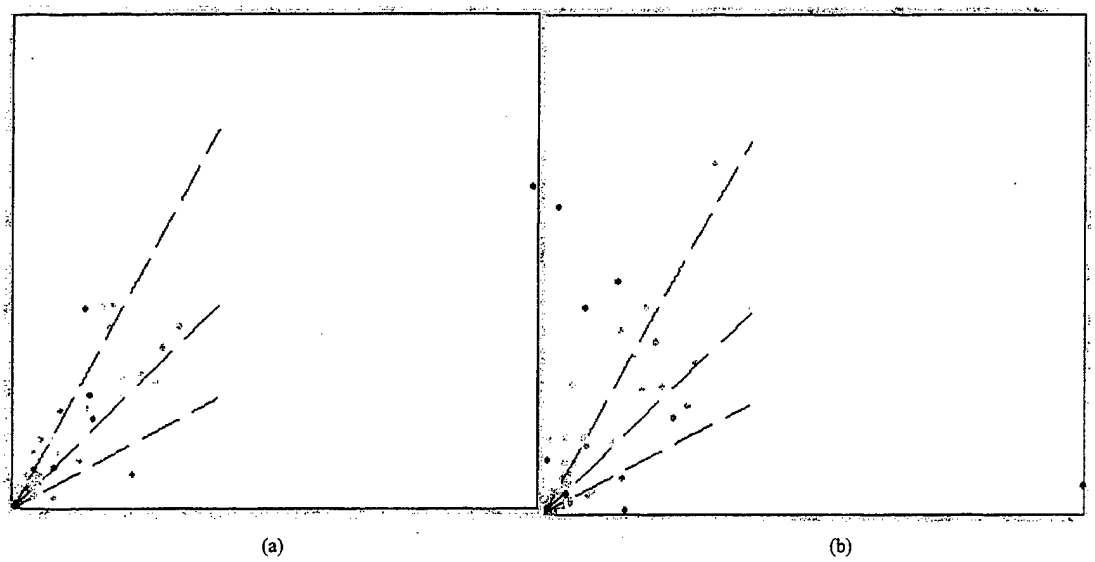


Fig. 4. Example of differential expression of housekeeping genes (red dots, where the dashed lines are the partition edges of the window functions). (a) *Within-class* and (b) *between-class*.

the expression of these housekeeping genes. For example, differential expression was observed between LCC1 and LCC9. See Fig. 4, where the data sets were normalized using the first method discussed above. This fact was observed from all of our experiments and shared by the same observation reported in [11]. Therefore, selection and use of housekeeping genes for normalization of differential expression data from various biological models should be approached with caution [11].

Since evaluation requires comparison with existing methods, we have implemented all three major approaches and applied these to the same data sets. In this experiment, all genes are considered as control genes and used in the calculation. Our measure of normalization accuracy is the MSE defined by (2) over the selected control gene set. The result of using the first method is given in Fig. 5, where $a = \sum_{i=1}^{n_c} y_i / \sum_{i=1}^{n_c} x_i = 9.9$ and $b = 0$; an MSE of 8549 is reached. In the second method, normalization is based on a linear regression through the origin, i.e., $a = \sum_{i=1}^{n_c} x_i y_i / \sum_{i=1}^{n_c} x_i^2$ that is most close to the correct formulation. The corresponding result is shown in Fig. 6, where $a = 5.0$ and $b = 0$. A lower MSE of 3905 is obtained, consistent with our theoretical expectation. In Fig. 7, we show the normalization result using the third method, i.e., $a = 1/n_c \sum_{i=1}^{n_c} (y_i/x_i)$. As predicted, a biased estimate of the expression ratio is obtained, leading to a high MSE of 20 728 with $a = 18$.

These comparisons clearly indicate that the three existing approaches are not equivalent, as shown by both our experimental results and the theoretical justification of (12). To illustrate the

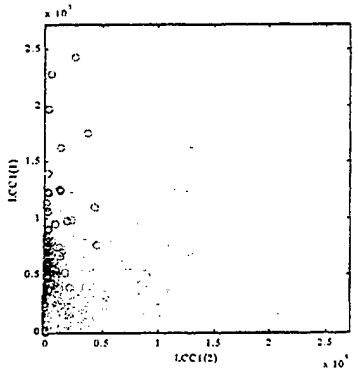


Fig. 5. Scatter plot of normalized microarray data using the existing method 1. (Circle: before normalization; dot: after normalization.)

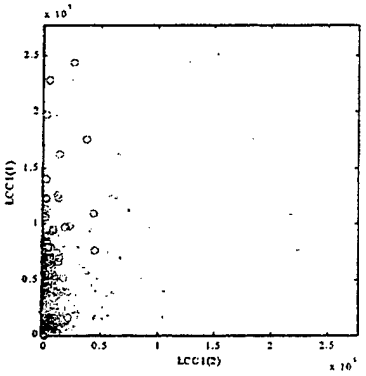


Fig. 6. Scatter plot of normalized microarray data using the existing method 2. (Circle: before normalization; dot: after normalization.)

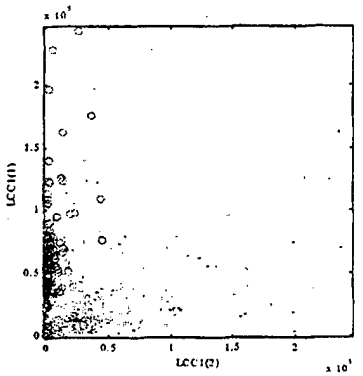


Fig. 7. Scatter plot of normalized microarray data using the existing method 3. (Circle: before normalization; dot: after normalization.)

impact of using the whole gene set as control gene set and using a dynamic programming technique on the normalization accuracy, we applied method 2 to the differential expression between LCC1 and LCC9. The scatter plot is given in Fig. 8. The corresponding MSE in this case is 6527, compared to the previous MSE of 3905. An increase in MSE suggests that, as samples become more divergent, a good normalization may be achieved using a subset of constantly expressed genes rather than a global normalization (e.g., using all genes) [3]. We then used the FS procedure to estimate both a and b. This additional step further

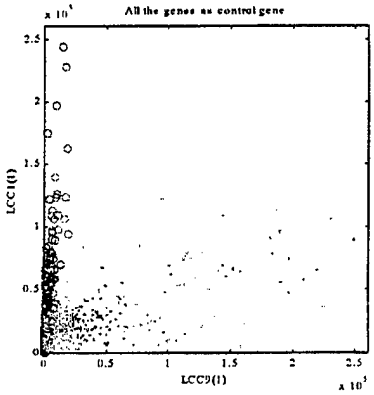


Fig. 8. Scatter plot of normalized *between-class* microarray data using the existing method 2. (Circle: before normalization; dot: after normalization.)

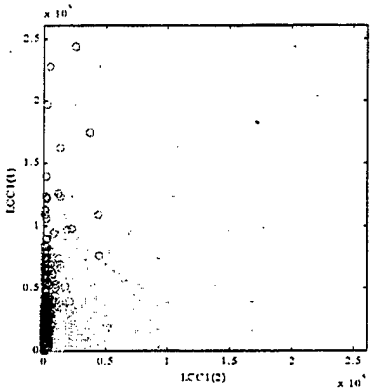


Fig. 9. Scatter plot of normalized *within-class* microarray data based on selected control gene subset using a static window function. (Circle: before normalization; dot: after normalization.)

reduced the MSE to 6438, but this reduction is probably not significant.

To explore the effect of control gene selection, we first performed an initial linear regression using the whole gene set. Four different window functions were configured to select control gene subsets where $\epsilon_2 \leq r \leq \epsilon_3$ and ϕ is the sector angle of the window function. Based on the selected control genes, we then applied a single linear regression to normalizing *within-class* samples. A numerical comparison on the normalization accuracy of using different control gene subsets is conducted, as reported in Table I. The main feature to note from these results is that, for different window functions, a stable estimate of the scaling factor a is obtained, while the shifting offset b varies significantly from case to case. In addition, the MSEs of normalizations in all three cases are comparable (i.e., 5632~5796). The scatter plot of the best normalization result is shown in Fig. 9.

We further applied the same procedure to processing *between-class* samples and observed similar data characteristics. The scaling factor in this case is about $a = 44$, while b varies substantially. Not surprisingly, an increase in MSE is observed (i.e., 6754~7384). Numerical analysis with different window functions shows the capable nature of the approach, since the interim estimate of linear regression coefficient is very stable with a satisfactory low MSE. Indeed, the robustness of

TABLE I
NUMERICAL COMPARISONS OF NORMALIZATION RESULTS BASED ON A DESIGNATED SUBSET OF CONTROL GENES WITH DIFFERENT WINDOW CONFIGURATIONS

Window($\times 10^3$)	$r \in (1,4), \phi = \frac{\pi}{16}$	$r \in (1,4), \phi = \frac{\pi}{8}$	$r \in (3,6), \phi = \frac{\pi}{8}$	$r \in (16,29), \phi = \frac{\pi}{8}$
Coefficient	$a = 7.6, b = 8311$	$a = 7.7, b = 2885$	$a = 7.7, b = 803$	$a = 7.7, b = -11378$
MSE	5633	5676	5693	5796

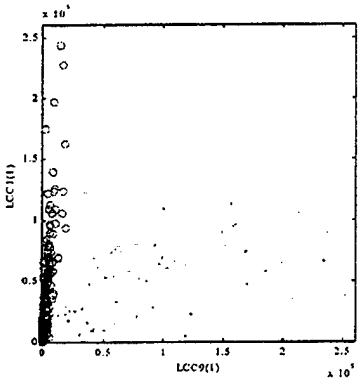


Fig. 10. Scatter plot of normalized *between-class* microarray data based on selected control gene subset using a static window function. (Circle: before normalization; dot: after normalization.)

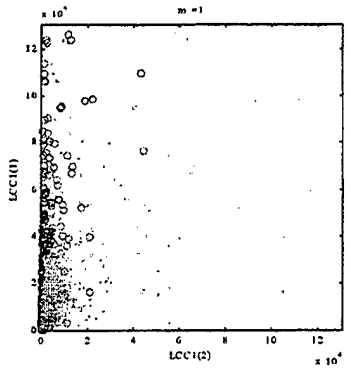


Fig. 11. Scatter plot of initial normalized *within-class* microarray data based on selected control gene subset using a dynamic window function. (Circle: before normalization; dot: after normalization.)

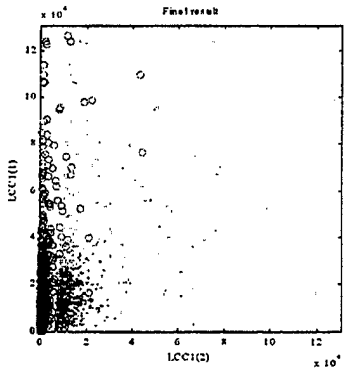


Fig. 12. Scatter plot of final normalized *within-class* microarray data based on selected control gene subset using a dynamic window function. (Circle: before normalization; dot: after normalization.)

the gene selection step has been successfully discovered in all our experiments. A typical scatter plot of *between-class* normalization results, using a window-based control gene selection, is given in Fig. 10.

Next, we provide an illustration of the iterative properties of our normalization algorithm. The sequence in our experiment shows the iterative recovery of the full linear regression matching. In this *within-class* case, $10\,000 < r < 24\,000$ and $\phi = \pi/2, \pi/4, \pi/8, \pi/16, \pi/32, \pi/48$. Each window shrinking step is mixed with one of the FS steps using the current set of recovered data points. The initial parameters are estimated based on the whole gene set. The normalization process converges to a good solution after six iterations. Figs. 11 and 12 show the scatter plots of initial and final normalization results. Once the algorithm has converged, the consistency of the control gene selection is significantly improved. Moreover, there are no erroneous matches between control genes for the last two adjacent iterations. The final control gene subset contains 37 genes. Finally, the MSE of 3892 is in good agreement with the corresponding results of the existing methods.

We next considered the iterative normalization for *between-class* samples. As a step toward improving the performance of microarray data normalization, we have put considerable effort into conducting various studies and developing reliable control gene selection and linear regression techniques. More precisely, we aim to perform an unsupervised normalization when confronted with unreliable housekeeping genes. Experience suggested that our newly proposed method can achieve this goal. We applied our algorithm to the differential expression between LCC1 and LCC9. In this *between-class* case, $10\,000 < r < 24\,000$ and $\phi = \pi/4, \pi/8, \pi/16, \pi/48$. As before, the initial parameters are estimated based on the whole gene set. The normalization process converges on a good solution after only four iterations. Figs. 13 and 14 show the scatter plots of initial and final normalization results. The final control gene subset contains 43 genes, and a stable and satisfactory MSE of 6523 is reached.

Finally, we used our previously developed the VISDA algorithm to display the expression patterns of different cell line samples in the gene expression space [13]. All data were normalized using the new method. For a molecular analysis of breast cancer, the profile of microarray expression is the molecular signature of interest. The representation of each sample is described as a point in a d-dimensional gene expression space in which each axis represents the expression level of one gene. The presence of well-separated sample groups implies that the representations of samples within the same group are close to each other in this gene expression space but distant from those of other samples. Thus, the representations of phenotype-specific samples form clusters. Fig. 15 shows a

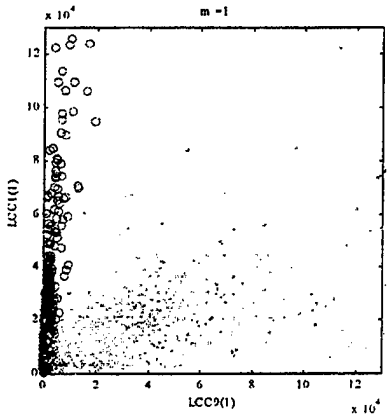


Fig. 13. Scatter plot of initial normalized *between-class* microarray data based on selected control gene subset using a dynamic window function. (Circle: before normalization; dot: after normalization.)

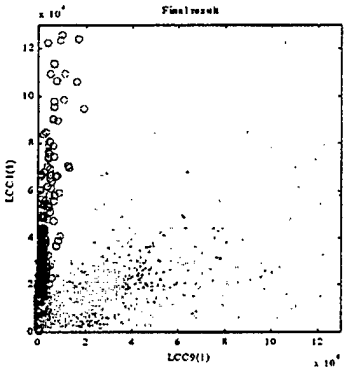


Fig. 14. Scatter plot of final normalized *between-class* microarray data based on selected control gene subset using a dynamic window function. (Circle: before normalization; dot: after normalization.)

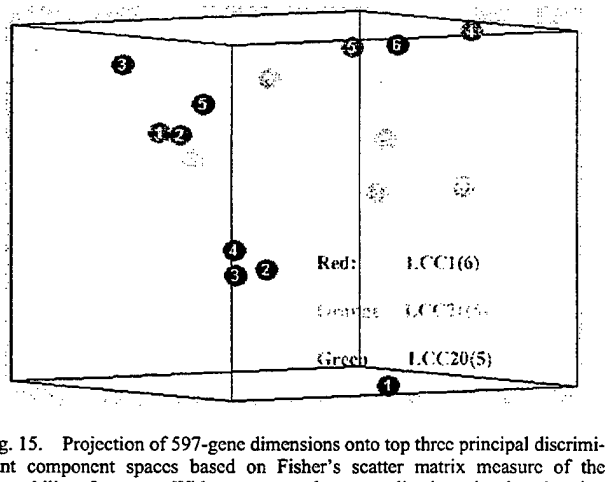


Fig. 15. Projection of 597-gene dimensions onto top three principal discriminant component spaces based on Fisher's scatter matrix measure of the separability of patterns. With an accurate data normalization, visual exploration reveals phenotype-specific sample clusters in gene expression space.

projected display of 597-gene dimensions into the top three principal discriminative component spaces, based on Fisher's scatter matrix [9]. With an accurate data normalization, visual exploration reveals three phenotype-specific sample clusters in gene expression space. Using the trace of Fisher's scatter matrix as a measure of the separability of patterns, our new normalization method achieved an improved performance with respect to the existing methods.

One important consideration with the present approach is the measure of quality in data normalization [11]. This is not a glamorous area, but progress in it is critical for the future success of data normalization [12]. What is the correct control gene set for a direct normalization of *between-class* data sets? How effective was a particular normalization method? Did the succeeding analysis come to the correct conclusion? Benchmark criteria assignment in data normalization are very different and difficult [5]. We believe that in data normalization, there is currently no objective measure of quality, and so it is difficult to quantify the merit of a particular data normalization technique. The effectiveness of such a techniques is often highly data-dependent. However, we would expect this iterative normalization method to be an effective tool in many gene microarray applications.

REFERENCES

[1] S. Kim, E. R. Dougherty, Y. Chen, K. Sivakumar, P. Meltzer, J. M. Trent, and M. L. Bittner, "Multivariate measurement of gene expression relationships," *Genomics*, vol. 67, pp. 201–209, 2000.

[2] A. J. Carlisle, V. V. Prabhu, A. Elkahloun, J. Hudson, J. M. Trent, W. M. Linehan, E. D. Williams, M. R. Emmert-Buck, L. A. Loitta, P. J. Munson, and D. B. Krizman, "Development of a prostate cDNA microarray and statistical gene expression analysis package," *Mol. Carcin.*, vol. 28, pp. 12–22, 2000.

[3] D. J. Duggan, M. L. Bittner, Y. Chen, P. Meltzer, and J. M. Trent, "Expression profiling using cDNA microarrays," *Nature*, vol. 21, pp. 10–14, January 1999.

[4] Y. Chen, E. R. Dougherty, and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *J. Biomed. Opt.*, vol. 2, pp. 364–374, 1997.

[5] P. Hedge, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. E. Hughes, E. Snesrud, N. Lee, and J. Quackenbush, "A concise guide to cDNA microarray analysis," *Biotechniques*, vol. 29, pp. 548–557, 2000.

[6] J. H. Zar, *Biostatistical Analysis*, 4th ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

[7] A. D. J. Cross and E. R. Hancock, "Graph matching with a dual-step EM algorithm," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 11, pp. 1236–1253, Nov. 1998.

[8] M. T. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar, "Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations," *Proc. Nat. Acad. Sci.*, vol. 97, no. 18, pp. 9834–9839, Aug. 2000.

[9] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

[10] A. P. Dhawan, L. K. Arata, A. V. Levy, and J. Mantil, "Iterative principal axes registration method for analysis of MR-PET brain images," *IEEE Trans. Biomed. Eng.*, vol. 42, pp. 1079–1088, Nov. 1995.

[11] S. T. Lott, "Array data normalization: Faith in housekeeping genes or regulation," in *3rd Annual Conf. Integrated Bioinformatics*, Zurich, Switzerland, Jan. 24–26, 2001.

[12] D. B. Finkelstein, J. Gollub, R. Ewing, F. Sterly, S. Somerville, and J. M. Cherry, "Iterative linear regression by sector: Renormalization of cDNA microarray data and cluster analysis weighted by cross homology," in *Critical Assessment of Techniques for Microarray Data Analysis*, Dec. 18–19, 2000.

[13] Y. Wang, L. Luo, M. T. Freedman, and S.-Y. Kung, "Probabilistic principal component subspaces: A hierarchical finite mixture model for data visualization," *IEEE Trans. Neural Nets*, vol. 11, pp. 625–636, May 2000.

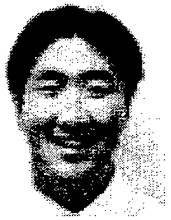
[14] R. Clarke, F. Leonessa, J. N. Welch, and T. C. Skaar, "Cellular and molecular pharmacology of antiestrogen action and resistance," *Pharmacol. Rev.*, vol. 53, pp. 25–71, 2001.

[15] Z. Gu, R. Lee, T. Skaar, J. Welch, K. Bouker, J. Lu, A. Liu, N. Davis, F. Leonessa, N. Brunner, Y. Wang, and R. Clarke, "Molecular profiles of antiestrogen resistance implicate NFkB, cAMP response element binding, nucleophosmin and interferon regulatory factor-1," J. Nat. Cancer Inst., 2001, submitted for publication.

Yue Wang received the Ph.D. degree in electrical engineering from the University of Maryland in 1995.

He is currently an Associate Professor of Electrical Engineering at The Catholic University of America, Washington, DC. He is also affiliated with the Johns Hopkins Medical Institutions as an Adjunct Professor of Radiology. His recent research focuses on computational bioinformatics and molecular imaging.

Jianping Lu is a Visiting Researcher at The Catholic University of America, Washington, DC, and an Associate Professor of Computer Science at SuZhou University, Su Zhou, China. His recent research focuses on bioinformatics, image processing, network, and database.



Richard Lee received the M.S. degree from the Department of Physiology and Biophysics, Georgetown University, Washington, DC, in 1994, where he is currently pursuing the Ph.D. degree.

His dissertation focuses on the study of retinoid resistance in breast cancer using computational bioinformatics.

Zhiping Gu received the Ph.D. degree in molecular biology from the University of Maryland in 1995.

She was with the Lombardi Cancer Center at Georgetown University during 1995–1999. She is currently a Research Scientist with Celera Genomics, Inc., MD. Her research interests focus on bioinformatics.

Robert Clarke received the Ph.D. and D.Sc. degrees in biochemistry from the Queen's University of Belfast, U.K., in 1986 and 1999, respectively.

He is currently a Professor in oncology, physiology, and biophysics at Georgetown University, Washington, DC. His research interests focus on studies into the molecular biology and endocrinology of breast cancer.

Dr. Clarke is a Fellow of the Royal Society of Chemistry (U.K.), Royal Society of Medicine (U.K.), and Royal Institute of Biology (U.K.).

[CANCER RESEARCH 62, 3428–3437, June 15, 2002]

Association of Interferon Regulatory Factor-1, Nucleophosmin, Nuclear Factor-κB, and Cyclic AMP Response Element Binding with Acquired Resistance to Faslodex (ICI 182,780)[1]

Zhiping Gu,[2] Richard Y. Lee, Todd C. Skaar,[3] Kerrie B. Bouker, James N. Welch, Jianping Lu, Aiyi Liu, Yuelin Zhu, Natalie Davis, Fabio Leonessa,[4] Nils Brünner,[5] Yue Wang, and Robert Clarke[6]

Vincent T. Lombardi Cancer Center and Department of Oncology, Georgetown University School of Medicine, Washington, DC 20007 [Z. G., R. Y. L., T. C. S., K. B. B., J. N. W., A. L., Y. Z., N. D., F. L., R. C.]; Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064 [J. L., Y. W.]; and Finsen Laboratory, Copenhagen, Denmark [N. B.]

ABSTRACT

To identify genes associated with survival from antiestrogens, both serial analysis of gene expression and gene expression microarrays were used to explore the transcriptomes of antiestrogen-responsive (MCF7/LCC1) and -resistant variants (MCF7/LCC9) of the MCF-7 human breast cancer cell line. Structure of the gene microarray expression data was visualized at the top level using a novel algorithm that derives the first three principal components, fitted to the antiestrogen-resistant and -responsive gene expression data, from Fisher's information matrix. The differential regulation of several candidate genes was confirmed. Functional studies of the basal expression and endocrine regulation of transcriptional activation of implicated transcription factors were studied using promoter-reporter assays.

The putative tumor suppressor interferon regulatory factor-1 is down-regulated in resistant cells, whereas its nucleolar phosphoprotein inhibitor nucleophosmin is up-regulated. Resistant cells also up-regulate the transcriptional activation of cyclic AMP response element (CRE) binding and nuclear factor κB (NFκB) while down-regulating epidermal growth factor receptor protein expression. Inhibition of NFκB activity by ICI 182,780 is lost in resistant cells, but CRE activity is not regulated by ICI 182,780 in either responsive or resistant cells. Parthenolide, a potent and specific inhibitor of NFκB, inhibits the anchorage-dependent proliferation of antiestrogen-resistant but not antiestrogen-responsive cells. This observation implies a greater reliance on their increased NFκB signaling for proliferation in cells that have survived prolonged exposure to ICI 182,780.

These data from serial analysis of gene expression and gene microarray studies implicate changes in a novel signaling pathway, involving interferon regulatory factor-1, nucleophosmin, NFκB, and CRE binding in cell survival after antiestrogen exposure. Cells can up-regulate some estrogen-responsive genes while concurrently losing the ability of antiestrogens to regulate their expression. Signaling pathways that are not regulated by estrogens also can be up-regulated. Thus, some breast cancer cells may survive antiestrogen treatment by bypassing specific growth inhibitory signals induced by antagonist-occupied estrogen receptors.

INTRODUCTION

ERs[7] are nuclear transcription factors, their activities being affected by the nature of the ligand bound and the pattern of genes/proteins expressed within cells (cellular context; Ref. 1). Antiestrogens compete with endogenous estrogens for activation of ER, and induce both cell cycle arrest and apoptosis in responsive cells (2). Neither the genes regulated by antiestrogens that signal to apoptosis nor those genes that confer an acquired antiestrogen resistance have been identified. Nonetheless, antiestrogenic drugs are effective in both premenopausal and postmenopausal breast cancer patients, and in the metastatic and adjuvant settings (3). The most widely used antiestrogen in current clinical practice is the triphenylethylene TAM. Clinical experience with this drug likely now exceeds 10 million patient years. When patients with metastatic disease are selected for treatment based on the ER and PgR content of their tumors, responses are seen in up to 75% of tumors expressing both receptors (2). TAM also reduces the incidence of ER-positive breast cancers in high risk women (4).

Other antiestrogens have emerged recently, most notably the benzothiophene Raloxifene and the steroidal ICI 182,780 (Faslodex). Both drugs appear to have significant clinical activity and may have better toxicological profiles when compared with TAM (2). Faslodex has significant activity in TAM-resistant patients (5), consistent with data obtained previously with TAM-resistant human breast cancer cells selected *in vitro* (6).

Despite the utility of antiestrogens, most tumors that initially respond to these drugs will recur and require alternative systemic therapies (2). Unfortunately, the precise mechanisms that confer resistance remain unknown. Change to an antiestrogen-stimulated phenotype has been described in some animal models (6, 7). This phenotype may occur in up to 20% of breast cancer patients but a loss of responsiveness to antiestrogens may be the more common phenotype (2). The expression of mutant ER proteins and splice variants has been reported but the functional role of these in endocrine resistance remains unclear (2). Most tumors acquiring antiestrogen resistance do so while retaining expression of ER (8). Thus, whereas lack of ER expression is a major form of *de novo* antiestrogen resistance, other mechanisms must be active in most instances of acquired resistance (2). The persistent expression of ER in tumors with acquired resistance suggests that some cells expressing this phenotype may either require ER expression and/or reflect the altered expression of otherwise estrogen-regulated genes.

Because ER-mediated transcription is directly affected by antiestrogens, we initially hypothesized that antiestrogen resistance might include perturbations in the patterns of expression and/or regulation of

Received 8/31/01; accepted 5/2/02.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] Supported in part by Public Health Service awards 5R01-CA/AG58022, 5P50-CA58185 (to R. C.), 5R33-CA83231 (to Y. W.) from the National Cancer Institute; Department of Defense Awards DAMD17-99-9189 (to K. B. B.), DAMD17-99-1-9191, BC010619, and BC990358 (to R. C.) from the United States Army Medical Research and Materiel Command; and the American Cancer Society IRG-97-1520-01 (to T. C. S.). Technical services also were provided by the Flow Cytometry and Cell Sorting, and Macromolecular Shared Resources funded through Public Health Service Award 2P30-CA51008 (Vincent T. Lombardi Cancer Center Support Grant).

[2] Present address: Celera Genomics, 45 West Gude Drive, Rockville, MD 20850.

[3] Present address: Indiana University, Department of Medicine, Indianapolis, IN 46202.

[4] Present address: Laboratory of Clinical Investigation, National Institute on Aging, NIH, 5600 Nathan Shock Drive, Baltimore, MD 21224.

[5] Present address: United States Patent and Trade Mark Office, Crystal Plaza 3, Washington, DC 20231.

[6] To whom requests for reprints should be addressed, at Room W405A Research Building, Vincent T. Lombardi Cancer Center, Georgetown University School of Medicine, 3970 Reservoir Road, NW, Washington, D.C. 20007. Phone: (202) 687-3755; Fax: (202) 687-7505; E-mail: clarker@georgetown.edu.

[7] The abbreviations used are: ER, estrogen receptor; CRE, cyclic AMP response element; CCS-IMEM, improved minimal essential medium supplemented with 5% charcoal calf stripped serum; EGF-R, epidermal growth factor receptor; IRF-1, interferon regulatory factor-1; NPM, nucleophosmin; PgR, progesterone receptor; SAGE, serial analysis of gene expression; TAM, Tamoxifen; XBP-1, X-box binding protein-1; FACS, fluorescence-activated cell sorting; NFκB, nuclear factor κB; EGR-1, early growth response factor-1; TNFα, tumor necrosis factor α.

a subset of all of the ER-regulated genes (1). To address this hypothesis, we first generated a novel series of human breast cancer variants from the MCF-7 human breast cancer cell line. These cells have different growth requirements for estrogen and exhibit differential sensitivities to TAM and ICI 182,780 (9–11). In this study, we focus on MCF7/LCC1 cells (estrogen-independent, TAM-responsive, and ICI 182,780 responsive) and MCF7/LCC9 cells (estrogen-independent, ICI 182,780 resistant, and TAM cross-resistant; Ref. 11). Because the cells exhibit comparable cell cycle profiles[8] and are both MCF-7 variants, we can exclude the altered expression of genes related solely to differences in both genetic background and cell cycle distribution. A direct comparison of these respective transcriptomes should identify genes associated with survival from long-term antiestrogen exposure.

Several techniques are now available to explore the transcriptomes of tumors and experimental models. However, the most effective approach remains a matter of debate (12). Studies in breast cancer have been limited, most simply attempting to identify the genes expressed in breast cancers. For example, a recent study by Perou *et al.* (13) explored data from excisional breast biopsies from 42 individuals. Gene clusters, identified by exploration of the data structure, include those associated with ER, HER-2, and IFN-induced genes. A similar cluster of IFN-regulated genes was identified in the breast cancer cell lines included in the NIH drug screening program (14). Studies comparing the gene expression profiles of specific breast cancer phenotypes include an examination of histologically different samples from a single breast cancer lesion (15) and a preliminary analysis of a TAM-stimulated xenograft model (16). None of these reports directly addressed either the function or potential role of the specific genes identified. We have used two different but complementary approaches, SAGE and gene expression microarrays. These approaches would not be expected to provide identical data because not all of the genes identified by SAGE are on the microarrays, some genes identified on the cDNA arrays may be confounded by cross-hybridization to homologous RNAs, and the ability to detect significant differences between the SAGE databases is affected by the relative abundance of the tags and the size of the databases. We approached both technologies as means to sample the transcriptomes of MCF7/LCC1 and MCF7/LCC9 cells, and to generate data that would allow us to begin testing our hypothesis implicating estrogen-regulated genes in antiestrogen resistance. We now show that cells can survive prolonged antiestrogen treatment by altering the expression, patterns of regulation, and functional activation of specific estrogen-regulated genes.

MATERIALS AND METHODS

Cell Lines. MCF7/LCC1 cells were derived from the estrogen-dependent MCF-7 human breast cancer cell line after selection for growth in ovariectomized nude mice (9, 17). MCF-7/LCC9 cells were obtained by an *in vitro* stepwise selection of the estrogen-independent but antiestrogen-responsive MCF7/LCC1 cells against the steroidal antiestrogen ICI 182,780 (Faslodex). MCF7/LCC9 cells are ICI 182,780 resistant and TAM cross-resistant, express ER and PgR, and exhibit an estrogen-independent but responsive phenotype (11). MCF7/LCC1 and MCF7/LCC9 cells were routinely passaged in Improved Minimal Essential Medium without phenol red (Biofluids, Bethesda, MD) supplemented with 5% CCS-IMEM. Serum was stripped of endogenous estrogens as described previously and is estimated to contain ≤10 fM estrogen (18). Vehicle for all of the hormone/antihormone treatments was ethanol (final concentration <0.1% v/v). All of the cell cultures were maintained at 37°C in a humidified 5% CO_2:95% air atmosphere and shown to be free of contamination with *Mycoplasma* species as determined by solution hybridization to *Mycoplasma*-specific, radiolabeled, RNase riboprobes (Gen-Probe Inc., San Diego, CA).

SAGE Analyses. SAGE was performed as described previously (19). Polyadenylic acid mRNA was harvested from cells using biotin labeled-oligodeoxythymidylic acid magnetic beads (Promega PolyATract System 1000 kit; Promega, Madison, WI) and treated with DNase I enzyme to remove any contaminating DNA. mRNA (5 μg) was converted to double-stranded cDNA using the Life Technologies, Inc. cDNA Synthesis kit (Life Technologies, Inc., Rockville, MD). Biotinylated cDNA was completely cleaved with Nia III and the 3′-end digested fragments extracted with magnetic streptavidin beads. The cDNA was evenly divided and ligated, one half to linker A and the other half to linker B (19). Cleavage of the cDNA by BsmF1 produced 11–13 bp oligo DNA tags with linkers, which were blunt-ended with T4 polymerase. Linkers A and B were ligated together to form ditags, which were then amplified by PCR using primers to linkers A and B. Ditags (22–26 bp) were gel purified and ligated into concatenated polytags. The polytags were purified and cloned into the Sph1-digested pZeor1 vector, which was transferred to competent TOP10F′ cells by electroporation. Positive clones were selected overnight at 37°C for growth on low-salt Luria-Bertani bacterial plates supplemented with Luria-Bertani-Zeocin (50 μg/ml) and isopropyl β-D-thiogalactopyranoside (1 mM). Colonies were screened for plasmids containing appropriate inserts by size fractionating PCR products, obtained using M13 forward and reverse primers, in agarose gels. PCR products containing concatamers of >600 bp were purified and sequenced.

Characteristics of the SAGE databases are shown in Table 1. We compared the MCF7/LCC1 and MCF7/LCC9 databases, using the SAGE version 1.00 software (kindly provided by Dr. K. W. Kinzler, Johns Hopkins University, Baltimore, MD), to identify putatively differentially expressed genes. Only a representative sample of these can be presented. The genes presented in Table 2 were primarily selected based on: (*a*) fold difference ≥2-fold; (*b*) that the Tags compared should represent ≤2 genes; and (*c*) that a Tag found in either the MCF7/LCC1 and/or MCF7/LCC9 SAGE libraries must represent ≥0.10% of the database. Evidence that a gene was already known to be expressed in breast cancers also was considered. None of these criteria were considered an absolute requirement for gene selection. Whereas 2-fold was selected as the cutoff, biologically critical events can be controlled by genes that exhibit a fold regulation as small as 50% (20). As described recently by Man *et al.* (21), χ^2 analyses were used to compare the proportions of specific tags in each database.

RNA Isolation, Generation of Probes, and Hybridization of Gene Microarrays. Each probe was generated from an independent cell culture, each culture being grown on a different day but using identical cell culture conditions. Six MCF7/LCC1 and five MCF7/LCC9 cell cultures were used. RNA was isolated from proliferating, subconfluent monolayers of each cell line using the TRIzol reagent (Life Technologies, Inc., Grand Island, NY). RNA quality was determined by standard spectroscopic and gel electrophoresis analyses.

Probes for the Clontech Atlas gene microarrays (Clontech, Palo Alto, CA) were made as described by the manufacturer. Briefly, 1 μg of Dnase-treated mRNA was primed with the Clontech cDNA Synthesis Primer mix. The product was reverse transcribed into radiolabeled cDNA with [γ-^{32}P]dATP (Amersham Life Science Inc., Arlington Heights, IL), and the reaction incubated at 50°C for 25 min and terminated by adding 0.1 M EDTA (pH 8.0). Radiolabeled cDNA was purified and eluted through a NucleoSpin Extraction Column (centrifuged at 14,000 rpm). The cDNA probe was denatured with 1

Table 1 *Characteristics of the SAGE libraries from MCF7/LCC1 and MCF7/LCC9 cells*

Characteristics of SAGE libraries		Tags[a]	Gene hits
Tags sequenced from MCF7/LCC1 cells	12,816[b]	5,783	1
Tags sequenced from MCF7/LCC9 cells	11,109[b]	1,170	2
Number of Tags identified	10,518	208	3
Number of known Tags[c]	7,221	38	4
Number of unknown Tags	3,297	10	5

[a] Number of Tags representing a corresponding number of gene hits, *e.g.*, 5,783 Tags are specific for single genes, whereas 208 Tags could identify up to 3 genes each.
[b] Number of Tags in each SAGE database.
[c] Includes expression sequence tags.

[8] R. Clarke, unpublished observations.

Table 2 *Differentially expressed genes identified in the MCF7/LCC1 and MCF7/LCC9 SAGE libraries*

Putative gene[a]	Unigene no.	MCF7/LCC1	MCF7/LCC9	Difference[b]	P[c]	Gene function
N-ras-related gene	Hs.260523	2	20	10-fold	<0.001	G-protein
Cathepsin D	Hs.343475	7	34	5-fold	<0.001	Protease involved in tumor invasion
XBP-1	Hs.149923	7	25	4-fold	<0.001	Transcription factor
Prefoldin 5	Hs.288856	6	21	4-fold	0.002	Chaperone for unfolded proteins
HSP-27	Hs.76067	23	55	2-fold	0.001	Stress response protein
Vit B-12-binding protein	Hs.2012	17	37	2-fold	0.002	Vitamin-binding protein
NPM	Hs.9614	10	14	1.5-fold	>0.05	Oncogenic nucleolar protein
L14	Hs.738	13	2	6-fold	0.021	Ribosomal protein
Death-associated protein-6	Hs.336916	11	2	6-fold	0.049	Apoptosis-associated protein
EF-γ	Hs.2186	22	6	4-fold	0.014	Translation elongation factor
Ferritin, heavy polypeptide-1	Hs.62954	54	16	3-fold	<0.001	Iron-binding protein

[a] The gene designations are considered putative, although the identity of most genes designated in this fashion have been shown to be correct. These genes include those Tags where: (*a*) the fold difference is ≥2-fold; (*b*) the Tag could represent ≤2 genes: and (*c*) represents 0.1% of either the MCF7/LCC1 and/or MCF7/LCC9 SAGE library.

[b] Predicted fold difference in gene expression between MCF7/LCC1 *vs.* MCF7/LCC9 cells.

[c] Obtained by χ^2 analyses; P estimated to 3 significant figures.

[d] NPM (not statistically significant) is shown because we know it to be both estrogen regulated and associated with TAM treatment in patients.

M NaOH and 10 mM EDTA, and incubated at 68°C for 20 min. c_0t-1 DNA and 1 M NaH_2PO_4 (pH 7.0) were added to the denatured probe, and incubated at 68°C for an additional 10 min.

Each Atlas Array (Clontech) was prehybridized with 5 ml of ExpressHyb buffer (Clontech) and 0.5 mg of denatured DNA from sheared salmon testes at 68°C for 30 min with continuous agitation. The cDNA probe, prepared as described above, was then added and allowed to hybridize overnight. The array was washed four times with 2× SSC containing 1% (w/v) SDS for 30 min at 68°C and once with 0.1× SSC containing 0.5% (w/v) SDS for 30 min at 68°C. One final wash was performed with 2× SSC for 5 min at room temperature. The Atlas Array was sealed in plastic and signals detected by phosphorimage analysis using a Molecular Dynamics Storm phosphorimager (Molecular Dynamics, Sunnyvale, CA). Each filter was used only once.

Measuring NPM and EGF-R Protein Levels. Established methods were used for performing and quantifying Western analyses of NPM (22, 23). Briefly, 10 μg of protein was loaded onto an SDS-PAGE gel and fractionated under reducing conditions [5% (v/v) β-mercaptoethanol]. To account for within-gel differences, samples were loaded in a random sequence onto each gel. Proteins were blotted onto nitrocellulose membrane and the blots probed with an anti-NPM monoclonal antibody (kindly provided by Dr. Pui-Kwong Chan, Baylor College of Medicine, Houston, TX; Ref. 24). After transfer to the membranes, equal protein loading was confirmed by staining the nitrocellulose with Ponceau S as is widely reported (22, 23, 25). Any material remaining in the gels were stained by Coomassie Blue. This approach provides an adequate and appropriate estimate for equivalence of protein loading (22, 23, 25). Immunoreactivity was visualized using a horseradish peroxidase-linked goat antimouse IgG and the enhanced chemiluminescence detection system (Amersham Life Science Inc.). Chemiluminescence was densitometrically measured using a Quantity One Scanning and Analysis System (pdi, Huntingdon, NY).

EGF-R is expressed at low levels in MCF-7 cells and cannot readily be detected/quantified by Western blot. Consequently, we measured immunofluorescently labeled EGF-R protein by FACS. For each cell line, EGF-R immunofluorescence was performed by rinsing 5×10^6 cells once in PBS and pelleting cells by centrifugation at 1000 rpm for 5 min at room temperature. Cell pellets were resuspended in 100 μl of an anti-EGF-R mouse monoclonal antibody that recognizes the extracellular domain of the receptor (EGF-R antibody-1; NeoMarkers, Lab Vision Corp., Fremont, CA; 200 μg/ml diluted 1:50 in PBS), and incubated at room temperature for 1 h. Cell pellets were then resuspended in 1:50 dilution of R-phycoerythrin-conjugated goat antimouse IgG-2a (CALTAG Laboratories, Burlingame, CA) and incubated in the dark for 30 min. After rinsing in PBS, cells were again pelleted, fixed by resuspending in 1% paraformaldehyde, and fluorescence measured by FACS. Control cells were treated either with secondary antibody alone or with no antibody. FACS was performed on a FACStar^{Plus} flow cytometer (Becton-Dickinson, Mountain View, CA) at 488 nm.

RNase Protection Analysis of IFN Regulatory Factor-1 mRNA Expression. Total RNA was isolated using the TRIzol reagent (Life Technologies, Inc.) according to the manufacturer's instructions. The IRF-1 riboprobe was made by *in vitro* transcription of a 360-bp fragment of the IRF-1 cDNA. The 36B4 loading control riboprobe was similarly obtained from a 220-bp fragment of the 36B4 cDNA (17). Riboprobes were labeled by the addition of [^{32}P]UTP (Amersham Life Sciences Inc.) in the transcription buffer. To achieve bands for the two genes with similar intensities, the 36B4 riboprobe was made with a specific activity of ~20% that of the IRF-1 riboprobe. The RNase protection assays were performed as described previously (26). Briefly, total RNA (30 μg), the IRF-1 riboprobe, and the 36B4 riboprobe were hybridized overnight at 50°C. After digestion with RNase A, the protected fragments were size fractionated on 6% acrylamide Tris-borate EDTA-urea minigels (Novex, San Diego, CA). The gels were dried and the respective signals quantified by phosphorimager analysis (Molecular Dynamics).

Estimation of the Transcriptional Activation of CREs and NFκB. Two commercially available promoter-reporter assays were used to measure NFκB and CRE transcriptional activities. Experiments were performed as described by the manufacturer (Stratagene, La Jolla, CA). Briefly, firefly luciferase reporter constructs, under the control of the appropriate enhancer elements and *trans*-activator constructs, were provided in the PathDetect *in vivo* signal transduction pathway *cis*-reporting system (Stratagene). Cells were grown to 90% confluence in 5% CCS-IMEM medium and seeded at 8×10^4 cells into each well of 24-well tissue culture dishes. After incubation for 12–24 h, cells were transiently transfected with the appropriate plasmids using the Qiagen Superfect transfection reagent as described by the manufacturer (Qiagen, Valencia, CA). The ratio of plasmid to Superfect reagent was 250 ng:1 μl, with a transfection time of 2.5 h.

Estrogen (5 nM) and ICI 182,780 treatments (10 nM) were administered for 48 h after transfection in CCS-IMEM. Transfected cells were harvested and firefly luciferase activity measured using the Stratagene assay system. Activity is expressed in relative light units from a 20-μl sample as detected by luminometry. Each measurement is from duplicate samples, independent experiments being repeated on different days. Normalization of transfection efficiency was made to the *Renilla* luciferase reporter construct, under the control of the cytomegalovirus promoter (Promega). The *Renilla* luciferase assay was performed using the Promega Dual-luciferase reporter assay system.

Assessment of Growth Response to Parthenolide. MCF7/LCC1 and MCF7/LCC9 cells were plated in 96-well tissue culture plates and incubated for 24 h in 0.2 ml of 5% CCS-IMEM. Medium was removed and replaced with fresh 5% CCS-IMEM containing either vehicle (0.1% DMSO) or parthenolide (300 nM and 600 nM). Cells were refed every third day with the appropriate cell culture medium. Cell growth was determined on day 6, using a crystal violet assay where dye uptake is directly related to cell number (27). Cells were incubated for 30 min with crystal violet stain [0.5% (w/v) crystal violet in 25% (v/v) methanol] at 25°C. Unincorporated stain was removed with deionized water and the cells allowed to dry at room temperature. Incorporated dye was extracted into 0.1 ml of 0.1 M sodium citrate in 50% (v/v) ethanol for 10–15 min at room temperature. Absorbance was read at 570 nm using a Molecular Devices V_{max} kinetic microplate reader.

Statistical Analyses and Analysis of Gene Expression Microarray Data. *t* tests were used to compare control and experimental groups as appropriate for the RNase protection, Western blot, promoter-reporter, and cell proliferation assays. All of the tests were two-tailed, with statistical significance established at $P \leq 0.05$, unless stated otherwise.

For the gene array studies, background signal was estimated locally and

subtracted from the signal obtained from its target cDNA, producing the background-corrected data. These corrections were done using the algorithms in Pathways 4.0 (Research Genetics Inc., Huntsville, AL). Background-corrected data were normalized to account for differences in probe-specific activity, hybridization, and other variables among replicates (28). Normalization was accomplished using the mean value of all of the background-corrected signals on each array.

Different approaches have been used to analyze data from gene array studies. Some methods are simply based on fold-regulation (29), others are more statistically based (16, 30), and/or apply an informatics-based exploration of data structure (31, 32). The optimal approach remains a subject of considerable debate (30). As with most gene microarray studies, our data set is high in dimensionality (597 dimensions) but the number of replicates is limited by the resource-intensive nature of the technology. The relatively few replicates limits the applicability of normal mixture models and other analyses that can operate in high dimensional data space (33, 34) and often generates noisy data sets.

Previously, we have reported a hierarchical visualization algorithm that can reveal all of the major aspects of the multimodal data points, which concurrently exist in a high dimensional gene expression space (35, 36). Using this algorithm, our data can be projected from 597 dimensions to two or three dimensions (multidimensional scaling). This is accomplished by respectively deriving the first three principal components fitted to the antiestrogen responsive (MCF7/LCC1) and resistant (MCF7/LCC9) gene expression data (Fig. 1). Thus, we evaluate the data structure subsets visually and assess whether these contain differentially expressed genes that may contribute to the respective phenotypes.

Because we can visualize data structure, our next priority was to identify a simple, supervised approach for reducing the dimensionality of the data without affecting its structure. Thus, we applied geometric and simple descriptive statistical approaches to the normalized data before and after a logarithmic transformation of these data. As noted previously, the distribution of the expression data for each gene is unknown (30), and it is unclear whether these violate the normal distribution required for parametric analyses. Indeed, it seems likely that the distribution assumption required will be normal for some genes and not for others. Whereas most investigators analyze data transformed by a logarithmic function, those genes with values that appear normally distributed before transformation may no longer have this distribution once transformed.

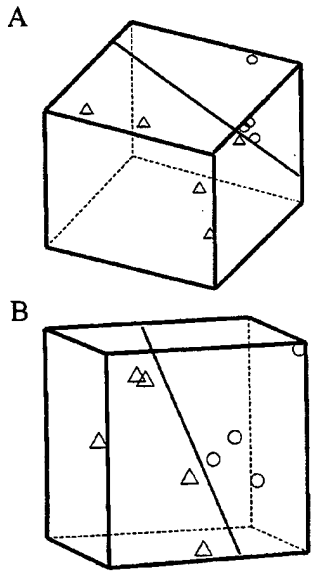


Fig. 1. Visual representations of the structure of the multidimensional gene microarray data. *A*, three-dimensional representation of 597 dimensions (△, MCF7/LCC1; ○, MCF7/LCC9) where the top three principal components capture 81.2% of the cumulative variance in the data. *B*, three-dimensional representation of 7 dimensions (data from Table 3) where the top three principal components capture 98.9% of the cumulative variance in the data. Axes represent the first three principal components derived from the gene expression data (79, 80). Plots are rotated to provide the optimal visualization. In both plots, a plane is shown demonstrating the linear separability of the MCF7/LCC1 (n = 5) and MCF7/LCC9 (n = 4) gene expression profiles.

To be inclusive, we used simple statistics (t tests) to explore the data. The inflated type-1 error from multiple comparisons should overestimate (false positive) significant differences. We considered this preferable to a high incidence of false-negative estimates, which would lead to the exclusion of potentially informative genes. The inclusion of uninformative genes (false negatives) is less problematic at this stage of the exploration. We used Student's t test, a t test for unequal variance (assumes normal distribution) and the nonparametric (distribution-free) Wilcoxon signed rank test. Logarithm transformed and nontransformed data were explored. This approach is similar to using a F test as described recently by Hedenfalk *et al.* (37).

t test results were evaluated and candidate genes selected with which to reconstruct a lower dimensional data set that should retain most of the information apparent in the top level visualization. However, the t test results were only one of several criteria used to guide gene selection, and only a subset of those genes that appear to be differentially regulated are presented. These genes were selected by comparing the results of t tests on logarithm transformed and untransformed data, fold-regulation (~2-fold or greater was selected because this difference is likely to be confirmed in independent analyses), the distribution of the background-corrected and normalized data for each gene (some genes appeared strongly differentially regulated but did not generate statistically significant differences because of heterogeneity in the data), and the probable relevance to breast cancer of each gene.

Where the gene subsets (reduced dimensional data) provide a reasonable description of the entire expression data, the replicate profiles of the resistant and responsive cells should exist in separable data space (35, 36). Furthermore, if the profiles are adequately defined by a small, rational gene subset, some of its members likely represent differentially expressed and functionally relevant genes. We acknowledge that our approach is limited, and is probably only applicable to simple comparisons within related cell culture models.

RESULTS

Genes Implicated by SAGE. The data in Table 1 show the number of different genes identified. Most genes were commonly expressed, and were not differentially expressed between the MCF7/LCC1 and MCF7/LCC9 cells. A selection of the genes identified by SAGE, and predicted to be differentially expressed in MCF7/LCC1 and MCF7/LCC9 SAGE databases, is shown in Table 2. Presentation of all of the genes expressed and/or differentially expressed is beyond the scope of a single, focused study.[9] The criteria applied for gene selection are described in "Materials and Methods." NPM was included because we already know it to be both estrogen regulated (23) and indirectly associated with TAM treatment in patients (38). Confirmation of the differential expression of NPM (see Table 2 and Fig. 2*B*) and altered CRE binding activity (the function of XBP-1; see Table 2 and Fig. 3*B*) indicate that these represent reasonable criteria for gene selection. Currently, the XBP-1 and NPM are the only genes from the SAGE database comparisons for which we have attempted to confirm differential expression/activation.

Comparing the SAGE databases identifies several genes that are up-regulated in MCF7/LCC9 cells compared with MCF7/LCC1 cells. These genes include *XBP-1, NPM, cathepsin D, HSP-27,* and n-*ras*. Increased CRE activity is indicated by the up-regulation of XBP-1, which regulates gene transcription through these response elements (39). XBP-1 is involved in regulating the expression of several tissue-specific genes including tissue inhibitor of metalloproteinases, osteopontin, and osteocalcin (40). Significantly, both Perou *et al.* (13) and West *et al.* (41) recently identified XBP-1 as being associated with ER gene expression clusters in human breast tumor biopsies. NPM is induced by estrogen in MCF-7 cells and is up-regulated in estrogen-independent cells (23). NPM also provokes an autoimmune

[9] http://clarkelabs.georgetown.edu/gu_et_al/gu_et_al_links.htm/.

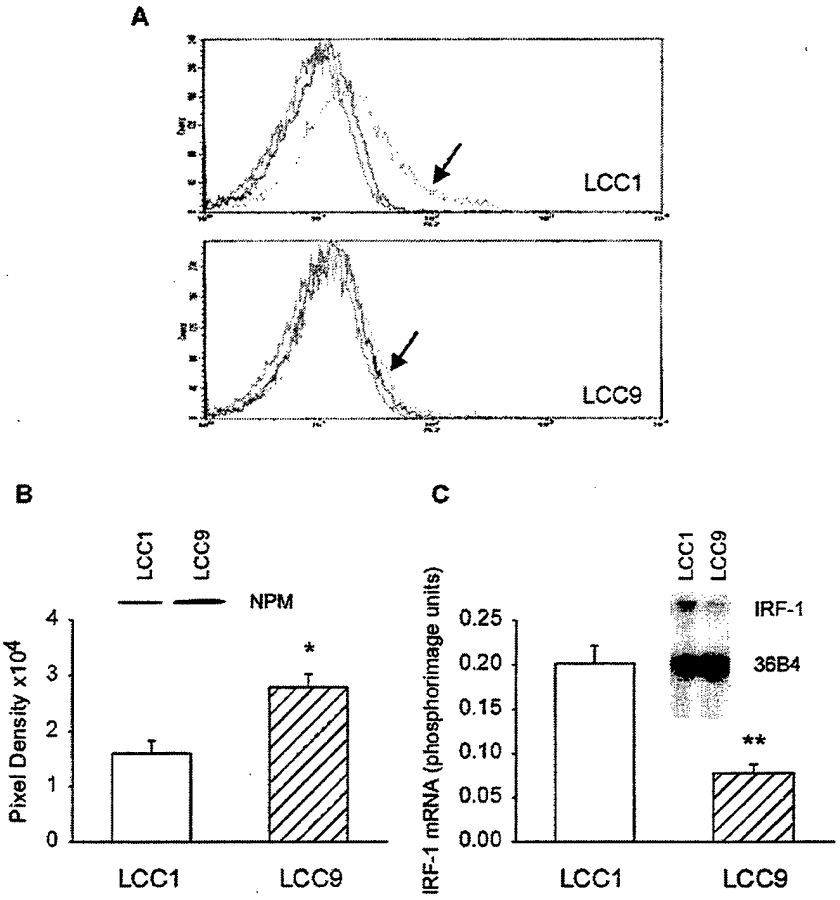


Fig. 2. Confirmation of the differential expression of NPM, EGF-R, and IRF-1 in MCF7/LCC1 and MCF7/LCC9 cells. *A*, EGF-R protein immunofluorescence as measured by FACS (representative figure of three experiments). *Arrows* indicate EGF-R signal, other signals are controls (no antibody; primary antibody but no secondary antibody). Axes are abscissa = fluorescence; ordinate = cell counts. *B*, NPM protein as measured by Western blotting ($*P \leq 0.02$) and represented as a percentage of control (MCF-7 cells growing in CCS-IMEM); *bars*, ±SE. Insert = representative Western blot. *C*, IRF-1 mRNA as measured by RNase protection ($**P = 0.005$, three independent replicate experiments) and expressed in phosphorimager units; *bars*, ±SE. Insert = representative analysis; 36B4 is a ribosomal gene (loading control).

response in breast cancer patients, the magnitude of which is associated with TAM therapy (38).

The altered expression of cathepsin D is consistent with our data published previously, showing increased secretion of this protein in several of our hormone-independent MCF-7 variants (42). Cathepsin D expression in breast tumors also is associated, at least in some studies, with a poor prognosis (43). HSP-27 expression has been implicated in refining the diagnosis of suspicious fine-needle aspirates of breast tissues (44). Vitamin B12 binding proteins are expressed in breast tumors (45), and vitamin B12 deficiency is a likely risk factor for breast cancer (46). Altered expression of the n-*ras*-related gene is consistent with the elevated *ras* signaling reported in some breast cancer cell lines and tumors (47).

SAGE also identified genes expressed at higher levels in the parental, antiestrogen-responsive cells (MCF7/LCC1) when compared with MCF7/LCC9 cells. These include ferritin, death-associated protein-6, and the eukaryotic elongation factor-γ. Ferritin is expressed in breast cancers, and breast tumor-derived ferritin may be a more useful tumor marker than measuring levels of ferritin in serum (48).

Structure of the Gene Microarray Data. It has been suggested that the cost required to perform gene microarray studies can be reduced by combining RNA populations from several replicates and performing a single hybridization on an Atlas array (16). However, we found heterogeneity among replicate experiments, which often remained after normalization. Logarithmic transformation of these data reduced this heterogeneity but not to the point where a single replicate could be used to obtain an adequate description of the data. Consequently, multiple replicates are required to provide a more reliable

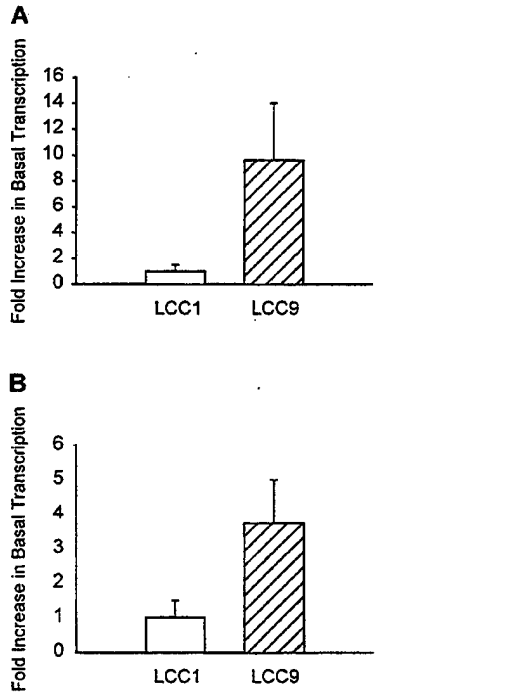


Fig. 3. Basal transcriptional activity of NFκB and CRE in MCF7/LCC1 and MCF7/LCC9 cells. *A*, NFκB. *B*, CRE. Data represent mean and are expressed as fold induction relative to MCF7/LCC1; *bars*, ±SE. All cells were grown in the absence of estrogens (CCS-IMEM).

Table 3 *Representative list of differentially expressed genes identified by gene microarray analyses*

Gene[a]	Unigene no.	MCF7/LCC1[b]	MCF7/LCC9	Gene function
NFκB	Hs.75569	1	2	Transcription factor involved in cell survival signaling
SOD	Hs.75428	1	2	Enzyme involved in detoxifying oxygen radicals
EGR-1	Hs.326035	3	1	Transcription factor
EGFR	Hs.77432	2	1	Growth factor receptor
IRF-1	Hs.80645	2	1	Transcription factor involved in signaling to cell cycle arrest and apoptosis
TNFα	Hs.241570	2	1	Cytokine
TNF-R1	Hs.159	2	1	Cytokine receptor involved in signaling to apoptosis

[a] Abbreviations are SOD, superoxide dismutase; TNF-R1, tumor necrosis factor-receptor 1.

[b] Data are represented as level of expression relative to the other cell line. Data are based on the mean values for each gene (6 microarrays of MCF7/LCC1; 5 microarrays of MCF7/LCC9). Values are expressed to the nearest integer.

estimate of the putative gene expression profiles. These observations on filter microarrays are consistent with recent reports for glass slide-based and oligonucleotide array-based gene expression microarrays (49, 50).

Fig. 1*A* is a visual representation of the multidimensional data (597 dimensions) in three dimensions. This visualization allows for an inspection of the data structure, and the likely comparability of the replicates among each other and between the two experimental groups (antiestrogen-responsive MCF7/LCC1 and antiestrogen-resistant MCF7/LCC9). For this top level visualization, the replicate gene expression profiles for MCF7/LCC1 and MCF7/LCC9 exist within linearly separable regions of the gene expression data space after elimination of one outlier array from each experimental group. The top three principal components capture 81.2% of the cumulative variance in the data (597 dimensions). Thus, the data structure is consistent with differences in the gene expression profiles as predicted by the known differential antiestrogen responsiveness of the two variants.

Genes Implicated by Gene Microarray Studies. The data in Table 3 show the fold-differences in expression of selected genes identified in the Clontech Atlas gene microarray studies selected using the criteria described in "Materials and Methods." The selection was not intended to describe fully the data set, only to assist in an initial exploration of the data. This small but rational subset of genes could be additionally evaluated in focused studies to confirm the differential expression patterns and establish potential functional relevance. Furthermore, if members of this subset were truly differentially expressed, we could begin to understand how cells perceive antiestrogens and adapt to this selective pressure.

To determine whether these genes are broadly representative of the differences between the gene expression profiles of MCF7/LCC1 and MCF7/LCC9 cells, we generated a three-dimensional projection from the seven-dimensional gene expression data space (Fig. 1*B*). This was necessary because we used several criteria to construct the subset, including some genes where fold-regulation or distribution of the data were given more weight than formal statistical significance. Consequently, we could not assume that we had maintained the linear separability of the data, at the top level, as seen in all 597 dimensions.

We might not expect this small subset of expression data (<2% of the information) to prove as effective in representing the respective phenotypes as the full data set (597 genes). Nonetheless, as for the 597-dimension visualization, after elimination of outlier data the seven-dimensional MCF7/LCC1 and MCF7/LCC9 profiles remain in linearly separable, three-dimensional data space. The top three principal components capture 98.9% of the cumulative variance in the data (seven-dimensions). This observation suggests that these data contain information that contributes to the differences in the molecular profiles of these two variants, that these genes may contribute to the respective biological phenotypes, and that additional studies of their potential functional relevance are warranted.

Genes expressed at a higher level in the MCF7/LCC1 cells include EGF-R, EGR-1, IRF-1, and both TNFα and its R1 receptor (TNF-R1). A well-established inverse relationship exists between the expression of EGF-R and ER in breast tumors (51). EGF-R can induce expression of EGR-1 (52), and expression of both genes is lower in MCF-7/LCC9 cells. EGR-1 is a transcription factor with proapoptotic activity (53) that can block NFκB function (54) and repress TGF-β receptor expression (29). EGR-1 expression is down-regulated in 7,12-dimethylbenz(*a*)anthracene-induced mammary adenocarcinomas in rats (55). IRF-1 is an IFN-regulated transcription factor that functions as a tumor suppressor gene (56, 57) and is induced by TNFα (58). A TNFα-mediated pathway for signaling to apoptosis occurs in MCF-7 human breast cancer cells (59, 60), and measuring serum TNF concentrations may be a useful prognostic marker in breast cancer patients (61). Furthermore, HER-2/*neu* can block resistance to TNFα-induced apoptosis in breast cancer cells, using a mechanism that involves activation of NFκB (62). We have previously implicated overexpression of superoxide dismutase in resistance to TNFα in MCF-7 cells (63). Superoxide dismutase appears to be up-regulated in MCF7/LCC9 cells (Table 3) and in TAM-stimulated MCF-7 xenografts (64). NFκB (p65/RelA) appears expressed at higher levels in MCF7/LCC9 cells. NFκB is overexpressed in ER-negative breast cancer cells (65) and has an important role in the development of the normal mammary gland (66).

NPM, EGF-R, and IRF-1 Are Differentially Expressed in MCF7/LCC1 and MCF7/LCC9 Cells. The data in Table 2 and Table 3 predict differential expression of NPM, EGF-R, and IRF-1 between MCF7/LCC1 and MCF7/LCC9 cells. To confirm these observations, we measured the levels of the EGF-R (immunofluorescence) and NPM proteins (Western blot) and IRF-1 mRNA (RNase protection). The data in Fig. 2*A* show that MFC7/LCC9 cells express lower amounts of EGF-R than MCF-7/LCC1 cells. NPM protein expression is significantly increased in MCF7/LCC9 cells compared with MCF7/LCC1 cells (Fig. 2*B*; $P < 0.02$), consistent with the predicted data from the SAGE analyses (Table 2) and our previous studies (23, 38). The higher levels of IRF-1 mRNA, seen in the antiestrogen-responsive MCF7/LCC1 cells in Table 3, are confirmed by RNase protection analysis (Fig. 2*C*; $P = 0.005$). Both the gene microarray and RNase protection analyses show an ~2-fold higher level of IRF-1 expression in MCF7/LCC1 cells, when compared with the antiestrogen-resistant MCF7/LCC9 cells.

Transcriptional Regulatory Activities of NFκB and CRE Are Increased in MCF7/LCC9 Cells. The increased expression of NFκB (gene expression microarray) and XBP-1 (SAGE) imply increased transcriptional activation of promoters containing NFκB and CRE response elements, respectively. We confirmed these observations directly, using commercially available promoter-reporter assays to measure transcriptional activities. The data in Fig. 3 show that the basal activity of both promoters is increased in MCF7/LCC9 cells; ~10-fold for NFκB and 4-fold for CRE ($P < 0.02$). The increase in transcriptional activation of the NFκB constructs is greater than that predicted by the gene array data, but mRNA, protein, and protein/DNA binding activities can be poor predictors of the functional activation of some transcription factors (67). This prediction is not problematic for XBP-1, where the 4-fold increase in mRNA expression identified by SAGE (Table 2) compares well with the 4-fold increase in basal transcriptional activation (Fig. 3*B*).

We next assessed whether ICI 182,780, the antiestrogen used to

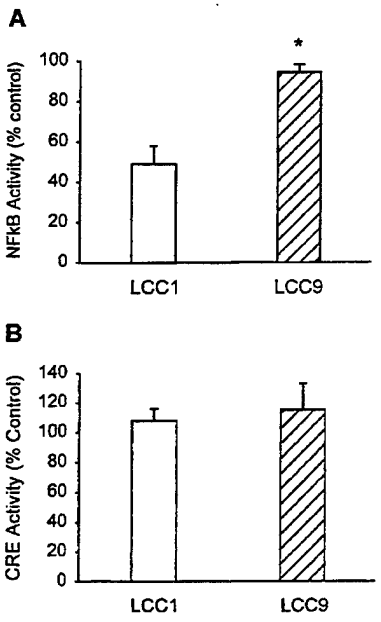


Fig. 4. Regulation of NFκB and CRE transcription by ICI 182,780 in MCF7/LCC1 and MCF7/LCC9 cells. *A*, NFκB (*$P \leq 0.001$, MCF7/LCC1 *versus* MCF7/LCC9). *B*, CRE (not significant). NFκB and CRE data are represented as mean of transcriptional activation expressed as a percentage of controls (vehicle-treated cells of the same cell line); *bars*, ±SE. Cells were grown in CCS-IMEM and treated with 10 nM ICI 182,780 for 48 h before measuring reporter gene expression.

generate the MCF7/LCC9 cells, could regulate the transcriptional activities of NFκB and CRE. Whereas ICI 182,780 inhibits NFκB activity in the MCF7/LCC1 cells (TAM- and ICI 182,780-responsive), this regulation is lost in the TAM and ICI 182,780 cross-resistant MCF7/LCC9 cells (Fig. 4*A*). In contrast, ICI 182,780 treatment does not alter the transcriptional regulatory activities of the CRE promoter in any of these variants (Fig. 4*B*).

MCF7/LCC9 Cells Are Specifically Responsive to an Inhibitor of NFκB Activity. The increased activation of NFκB and loss of its estrogenic regulation in MCF7/LCC9 cells suggests that these cells might now be partly dependent on NFκB signaling for survival/growth. Consequently, we compared the growth response of MCF7/LCC1 and MCF7/LCC9 cells to parthenolide, a potent and specific inhibitor of NFκB that can inhibit the inhibitor of NFκB kinase repressor of NFκB (68, 69) and also binds NFκB in a highly stereospecific manner to block DNA binding (70). Parthenolide produces a dose-dependent inhibition of MCF7/LCC9 cells, with an apparent IC_{50} of ~600 nM (Fig. 5). In contrast, parthenolide does not significantly affect growth of MCF7/LCC1 cells at these concentrations. MCF7/LCC9 cells are significantly more dependent on the transcriptional regulatory activities of NFκB than their ICI 182,780-responsive parental cells ($P < 0.01$ for MCF7/LCC9 *versus* MCF7/LCC1 at both 300 nM and 600 nM parthenolide).

DISCUSSION

We have begun to identify the molecular changes associated with cell survival after prolonged ICI 182,780 treatment in breast cancer cells. Whereas we have not attempted to confirm the altered expression of all implicated genes, some expression patterns are consistent with the activities we have confirmed. Here we discuss only those genes for which altered mRNA, protein, and/or transcriptional activation have been confirmed, and that are known to interact with each other in various cellular models, *i.e.*, IRF-1, NPM, NFκB, and CRE.

IRF-1 can function as a tumor suppressor and can signal to apoptosis through both p53-dependent and p53-independent pathways (71). These observations may partly reflect the ability of IRF-1 to induce a caspase cascade through activation of either caspase 1 (ICE; Ref. 72) and/or caspase 7 (73). Caspase 1 is involved in regulating apoptosis in normal mammary epithelial cells (74), and overexpression of caspase 1 is lethal in MCF-7 human breast cancer cells (75). Preliminary data from our laboratory demonstrate that overexpression of IRF-1 inhibits anchorage-dependent colony formation and that the rate of cell proliferation in MCF-7 cells is inversely related to the level of IRF-1 expression (76). These data suggest that the down-regulation of IRF-1 in MCF7/LCC9 cells may protect these cells from IRF-1-induced inhibition of proliferation and/or induction of apoptosis.

NPM can function as an oncogene, its overexpression fully transforming NIH 3T3 cells in a standard assay for oncogenic potential (77). We have shown that levels of autoantibodies to NPM increase in breast cancer patients 6 months before their recurrence. Consistent with an estrogenic/antiestrogenic regulation of NPM, the levels of these autoantibodies are lower in breast cancer patients that have received TAM (38). The increased NPM expression in MCF7/LCC9 cells compared with MCF7/LCC1 cells may reflect oncogenic potential of NPM, an activity potentially related to its ability to inhibit IRF-1 function (see below).

NFκB has been implicated in resistance to cytotoxic drugs and can function as a survival factor in various cell types (78). Several aspects of normal mammary gland development appear dependent on NFκB activity (66), perhaps partly reflecting its estrogenic regulation (65). Elevated NFκB activity arises early during neoplastic transformation in the rat mammary gland (79). Widely expressed in breast cancer cells and tumors, elevated NFκB activity is associated with estrogen-independence (65, 66). Currently, NFκB is the only protein known to induce BRCA2 expression (80). ICI 182,780 cannot suppress the increased NFκB activity in MCF7/LCC9 cells, despite inhibiting this function in ICI 182,780-responsive cells (MCF7/LCC1). The functional relevance of this observation was tested directly using parthenolide, which both specifically binds NFκB and blocks degradation of the endogenous NFκB inhibitor IκB, resulting in the inhibition of NFκB transcriptional regulatory activities (68, 70). This activity of parthenolide has been used to evaluate the functional role of NFκB in several recent studies (68, 69, 81, 82). MCF7/LCC9 cells are significantly more sensitive to growth inhibition by parthenolide than their MCF7/LCC1 parental cells. This

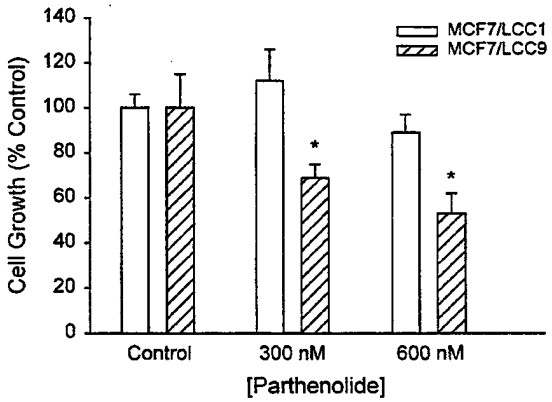


Fig. 5. Response to inhibition of NFκB activity by parthenolide. Data represent mean of four determinations, where absorbance in each treated population is expressed as a percentage of the absorbance in control cells (vehicle treated cells of the same cell line). *$P \leq 0.01$ MCF-7/LCC1 *versus* MCF7/LCC9. Cells were grown in CCS-IMEM without (control; vehicle only) or with parthenolide supplementation (300 nM; 600 nM).

observation is consistent with a greater functional reliance on NFκB activation for cell growth/survival, and implies that one option for surviving antiestrogen exposure is the up-regulation of an estrogen-regulated survival factor(s) concurrent with the loss of its ER-mediated regulation. Furthermore, parthenolide is now in clinical trials, and our data suggest that it may prove useful in combination with Faslodex or other antiestrogens to either increase responsiveness and/or delay the appearance of resistant disease.

XBP-1 has been identified recently in clusters of genes associated with ER-positive breast tumors in two independent studies (13, 41), and its expression is increased in MCF7/LCC9 cells. XBP-1 is a transcription factor that binds and activates CRE (39). The importance of CRE-regulated events is widely reported in many cell types (83, 84). These events include a likely role in signal transduction either at or downstream of ER and PgR (85). The relevance of increased CRE activity in MCF7/LCC9 cells is additionally supported by recent evidence that CRE-decoy oligonucleotides inhibit the growth of MCF-7 cells (86). We detected a 4-fold increase in CRE transcriptional activation in MCF7/LCC9 cells. Importantly, ICI 182,780 cannot regulate CRE activity in either MCF7/LCC1 (ICI 182,780-responsive) or MCF7/LCC9 (resistant) cells. These data imply an additional option available to breast cancer cells, a switch to signaling pathways that are normally independent of ER-mediated signaling.

IRF-1, NPM, NFκB, and CRE are known to affect cell proliferation, apoptosis, and/or carcinogenesis. Two critical protein-protein interactions directly link the IRF-1, NFκB, and NPM proteins. Direct binding occurs between IRF-1 and NPM (77), and between IRF-1 and NFκB (87, 88). In both cases, the interactions with IRF-1 have important effects on gene transcription and cell signaling. NPM binding inhibits the transcription regulatory activities of IRF-1 (77). A coordinated perturbation in the regulation of these two genes has occurred in the MCF7/LCC9 cells; NPM is up-regulated and IRF-1 is down-regulated. Thus, overexpression of NPM could additionally reduce the remaining lower levels of IRF-1, potentially blocking/eliminating its ability to initiate an apoptotic caspase cascade through caspase 1 and/or caspase 7. Such an effect would likely also eliminate the ability of IRF-1 to induce $p21^{cip1/waf1}$ (89) and cooperate with wild-type p53 in signaling to apoptosis (56, 57). Changes in the amount of available IRF-1 will directly affect the number of IRF-1:NFκB heterodimers available to regulate an additional series of genes. Whereas NFκB will compete with NPM for IRF-1 binding, their relative affinities for IRF-1 are unknown, and the preferred IRF-1 heterodimer remains to be established. IRF-1:NFκB protein-protein interactions or other cooperative interactions are implicated in the induction of ATF-2/jun (90), RANTES (91), VCAM-1 (88), interleukin 6 (92), and MHC class 1 genes (87). A functional IFN-β enhanceosome has been described that includes IRF-1, NFκB, and ATF2/jun (93). The importance of both IRF-1 and NFκB in IFN-induced signaling may contribute to the ability of IFNs to increase responses to antiestrogens (94–96).

CRE activation also may interact with the pathways regulated by IRF-1, NFκB, and NPM interactions. Delgado *et al.* (97) described a cyclic AMP-dependent pathway that inhibits IRF-1 transactivation. Thus, the increased CRE activity in MCF7/LCC9 cells may explain, in part, the lower IRF-1 mRNA levels seen both in the gene expression arrays and in the IRF-1 RNase protection studies.

The concurrent changes in NPM, IRF-1, NFκB, and CRE suggest a novel integrated signaling pathway that may involve the ability of NPM and CRE to inhibit IRF-1 initiation of a caspase cascade to apoptosis, the altered ability of cells to induce genes dependent on IRF-1:NFκB, and an increased activation of survival pathways that involve both NFκB and CRE. Studies to additionally establish the nature, function, and regulation of this putative pathway are currently in progress, including an overexpression of NFκB in sensitive cells and a dominant-negative approach in resistance cells. Because we looked only at cells that survived long-term antiestrogen exposure, the ability of the changes implicated in the present study to protect from an initial or short term exposure have yet to be determined. For example, cells may or may not survive an initial antiestrogenic exposure by the same mechanisms that allow for long-term survival. Irrespective of whether these other genes are functionally involved, their patterns of expression may be important in better predicting the 25% of ER+/PgR+, 55% of ER-/PgR+, and 66% of ER+/PgR− breast tumors that do not respond to antiestrogens (2).

It is not possible, in a single focused study, to define all of the potentially differentially expressed genes nor to establish their functional relevance firmly. Because the number of cellular models studied is small, additional functional studies where expression of the candidate genes is induced or repressed are in progress. Nonetheless, our data imply that breast cancer cells have highly plastic transcriptomes, with access to several signal transduction pathways for regulating the choice to differentiate, proliferate, or die. For example, MCF7/LCC9 cells have taken several possible interactive/interdependent approaches to circumvent the growth inhibitory effects of antiestrogens. This plasticity in gene expression patterns is consistent with the marked heterogeneity apparent in the clinical disease (2, 98).

In summary, our data suggest that one molecular profile associated with surviving prolonged antiestrogen exposure may include loss of ER-mediated signaling to apoptosis through IRF-1. This lost signaling is achieved both by down-regulation of IRF-1 and a coordinated up-regulation of its inhibitor NPM, and possibly another protein partner NFκB. Up-regulation of CRE activities also is implicated in this molecular profile. Other patterns of gene expression may provide alternative routes to the resistant phenotype or in cells that acquire a TAM-stimulated phenotype (2). The identification of these molecular profiles and signaling pathways may ultimately allow us to understand ER-regulated signaling, facilitate the development of novel treatment strategies, and allow clinicians to better identify antiestrogen-responsive and -resistant breast tumors.

ACKNOWLEDGMENTS

We thank Dr. K.W. Kinzler and his colleagues at Johns Hopkins University, Baltimore, MD, for their assistance in establishing the SAGE protocols and for providing their SAGE data analysis software.

REFERENCES

1. Clarke, R., and Brünner, N. Acquired estrogen independence and antiestrogen resistance in breast cancer: estrogen receptor-driven phenotypes?. Trends Endocrinol. Metab., *7:* 25–35, 1996.
2. Clarke, R., Leonessa, F., Welch, J. N., and Skaar, T. C. Cellular and molecular pharmacology of antiestrogen action and resistance. Pharmacol. Rev., *53:* 25–71, 2001.
3. Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomized trials. Lancet, *351:* 1451–1467, 1998.
4. Fisher, B., Costantino, J. P., Wickerham, D. L., Redmond, C. K., Kavanah, M., Cronin, W. M., Vogel, V., Robidoux, A., Dimitrov, M., Atkins, J., Daly, M., Wieand, S., Tan-Chiu, E., Ford, L., and Wolmark, N. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 study. J. Natl. Cancer Inst., *90:* 1371–1388, 1998.
5. Howell, A., DeFriend, D. J., Robertson, J. F. R., Blamey, R. W., Anderson, L., Anderson, E., Sutcliffe, F. A., and Walton, P. Pharmacokinetics, pharmacological and anti-tumor effects of the specific anti-oestrogen ICI 182780 in women with advanced breast cancer. Br. J. Cancer, *74:* 300–308, 1996.
6. Gottardis, M. M., and Jordan, V. C. Development of tamoxifen-stimulated growth of MCF-7 tumors in athymic mice after long-term antiestrogen administration. Cancer Res., *48:* 5183–5187, 1988.
7. Osborne, C. K., Coronado, E. B., and Robinson, J. P. Human breast cancer in athymic nude mice: cytostatic effects of long-term antiestrogen therapy. Eur. J. Cancer Clin. Oncol., *23:* 1189–1196, 1987.

8. Johnston, S. R. D., Saccani-Jotti, G., Smith, I. E., Newby, J., and Dowsett, M. Change in oestrogen receptor expression and function in tamoxifen-resistant breast cancer. Endocr. Related Cancer, *2:* 105–110, 1995.
9. Clarke, R., Brünner, N., Katzenellenbogen, B. S., Thompson, E. W., Norman, M. J., Koppi, C., Paik, S., Lippman, M. E., and Dickson, R. B. Progression from hormone dependent to hormone independent growth in MCF-7 human breast cancer cells. Proc. Natl. Acad. Sci. USA, *86:* 3649–3653, 1989.
10. Brünner, N., Frandsen, T. L., Holst-Hansen, C., Bei, M., Thompson, E. W., Wakeling, A. E., Lippman, M. E., and Clarke, R. MCF7/LCC2: A 4-hydroxytamoxifen resistant human breast cancer variant which retains sensitivity to the steroidal antiestrogen ICI 182, 780. Cancer Res., *53:* 3229–3232, 1993.
11. Brünner, N., Boysen, B., Jirus, S., Skaar, T. C., Holst-Hansen, C., Lippman, J., Frandsen, T., Spang-Thomsen, M., Fuqua, S. A. W., and Clarke, R. MCF7/LCC9: an antiestrogen resistant MCF-7 variant in which acquired resistance to the steroidal antiestrogen ICI 182, 780 confers an early cross-resistance to the non-steroidal antiestrogen tamoxifen. Cancer Res., *57:* 3486–3493, 1997.
12. Marx, J. DNA arrays reveal cancer in its many forms. Science (Wash. DC), *289:* 1670–1672, 2000.
13. Perou, C. M., Sorlie, T., Eisen, M. B., Van de, R. M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Borresen-Dale, A. L., Brown, P. O., and Botstein, D. Molecular portraits of human breast tumours. Nature (Lond.), *406:* 747–752, 2000.
14. Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de, R. M., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. Systematic variation in gene expression patterns in human cancer cell lines. Nat. Genet., *24:* 227–235, 2000.
15. Sgroi, D. C., Teng, S., Robinson, G., LeVangie, R., Hudson, J. R., and Elkahloun, A. G. *In vivo* gene expression profile analysis of human breast cancer progression. Cancer Res., *59:* 5656–5661, 1999.
16. Hilsenbeck, S. G., Friedrichs, W. E., Schiff, R., O'Connell, P., Hansen, R. K., Osborne, C. K., and Fuqua, S. A. W. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. J. Natl. Cancer Inst., *91:* 453–459, 1999.
17. Brünner, N., Boulay, V., Fojo, A., Freter, C., Lippman, M. E., and Clarke, R. Acquisition of hormone-independent growth in MCF-7 cells is accompanied by increased expression of estrogen-regulated genes but without detectable DNA amplifications. Cancer Res., *53:* 283–290, 1993.
18. Darbre, P., Yates, J., Curtis, S., and King, R. J. B. Effect of estradiol on human breast cancer cells in culture. Cancer Res., *43:* 349–354, 1983.
19. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. Serial analysis of gene expression. Science (Wash. DC), *270:* 484–487, 1995.
20. Yan, H., Dobbie, Z., Gruber, S. B., Markowitz, S., Romans, K., Giardello, F. M., Kinzler, K. W., and Vogelstein, B. Small changes in expression affect predisposition to tumorigenesis. Nat. Genet., *30:* 25–26, 2002.
21. Man, M. Z., Wang, X., and Wang, Y. POWER SAGE: comparing statistical tests for SAGE experiments. Bioinformatics (Oxf.), *16:* 953–959, 2000.
22. Harlow, E., and Lane, D. (eds.). Antibodies. A Laboratory Manual. Cold Spring Harbor, NY: CSH, 1988.
23. Skaar, T. C., Prasad, S. C., Sharach, S., Lippman, M. E., Brünner, N., and Clarke, R. Two-dimensional gel electrophoresis analyses identify nucleophosmin as an estrogen regulated protein associated with acquired estrogen-independence in human breast cancer cells. J. Steroid Biochem. Mol. Biol., *67:* 391–402. 1998.
24. Chan, P. K., Chan, W-Y., Yung, B. Y. M., Cook, R. G., Aldrich, M., Ku, D., Goldknopf, I. L., and Busch, H. Amino acid sequence of a specific antigenic peptide of protein B23. J. Biol. Chem., *261:* 14335–14341, 1986.
25. Klein, D., Kern, R. M., and Sokol, R. Z. A method for quantification and correction of proteins after transfer to immobilization membranes. Biochem. Mol. Biol. Int., *36:* 59–66, 1995.
26. Clarke, R., Brünner, N., Katz, D., Glanz, P., Dickson, R. B., Lippman, M. E., and Kern, F. The effects of a constitutive production of TGF-α on the growth of MCF-7 human breast cancer cells *in vitro* and *in vivo*. Mol. Endocrinol., *3:* 372–380, 1989.
27. Frandsen, T. L., Boysen, B. E., Jirus, S., Spang-Thomsen, M., Danø, K., Thompson, E. W., and Brünner, N. Experimental models for the study of human cancer cell invasion and metastasis. Fibrinolysis, *6*(Suppl. 4)*:* 71–76, 1992.
28. Wang, Y., Lu, J., Lee, R. Y., and Clarke, R. Iterative normalization of cDNA microarray data. IEEE Trans. Inf. Technol. Biomed., *6:* 29–36, 2002.
29. Du, B., Fu, C., Kent, K. C., Bush, H., Jr., Schulick, A. H., Kreiger, K., Collins, T., and McCaffrey, T. A. Elevated Egr-1 in human atherosclerotic cells transcriptionally represses the transforming growth factor-β type II receptor. J. Biol. Chem., *275:* 39039–39047, 2000.
30. Wittes, J., and Friedman, H. P. Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data. J. Natl. Cancer Inst., *91:* 400–401, 1999.
31. Ermolaeva, O., Rastogi, M., Pruitt, K. D., Schuler, G. D., Bittner, M. L., Chen, Y., Simon, R., Meltzer, P., Trent, J. M., and Boguski, M. S. Data management and analysis for gene expression arrays. Nat. Genet., *20:* 19–23, 1998.
32. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (Wash. DC), *286:* 531–537, 1999.
33. Hinneburg, A., and Keim, D. A. Optimal grid-clustering: towards breaking the curse of dimensionality in high-dimensional clustering. *In* M. P. Atkinson, M. E. Orlowska, P. Valduriez, S. B. Zdonik, and M. L. Brodie (eds.), Proceedings of the 25th Conference on Very Large Databases, pp. 506–517. San Francisco: Morgan Kaufman, 1999.
34. Lu, J., Wang, Y., Xuan, J., Kung, S. Y., Gu, Z., and Clarke, R. Discriminative mining of gene microarray data. Proc. IEEE Neural Netw. Signal. Process., *11:* 218–227, 2001.
35. Wang, Y., Luo, L., Freedman, M. T., and Kung, S. Y. Probabilistic principal component subspaces: a hierarchical finite mixture model for data visualization. IEEE Trans. Neural Netw., *11:* 635–646, 2000.
36. Wang, Y., Lin, S. H., Li, H., and Kung, S. Y. Data mapping by probabilistic modular networks and information theoretic criteria. IEEE Trans. Signal Process., *46:* 3378–3397, 1998.
37. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., and Trent, J. Gene-expression profiles in hereditary breast cancer. N. Engl. J. Med., *344:* 539–548, 2001.
38. Brankin, B., Skaar, T. C., Trock, B. J., Berris, M., and Clarke, R. Autoantibodies to numatrin: an early predictor for relapse in breast cancer. Cancer Epidemiol. Biomark. Prev., *7:* 1109–1115, 1998.
39. Clauss, I. M., Chu, M., Zhao, J. L., and Glimcher, L. H. The basic domain/leucine zipper protein hXBP-1 preferentially binds to and transactivates CRE-like sequences containing an ACGT core. Nucleic Acids Res., *24:* 1855–1864, 1996.
40. Clauss, I. M., Gravallese, E. M., Darling, J. M., Shapiro, F., Glimcher, M. J., and Glimcher, L. H. *In situ* hybridization studies suggest a role for the basic region-leucine zipper protein hXBP-1 in exocrine gland and skeletal development during mouse embryogenesis. Dev. Dyn., *197:* 146–156, 1993.
41. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Jr., Marks, J. R., and Nevins, J. R. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. USA, *98:* 11462–11467, 2001.
42. Thompson, E. W., Brünner, N., Torri, J., Johnson, M. D., Boulay, V., Wright, A., Lippman, M. E., Steeg, P. S., and Clarke, R. The invasive and metastatic properties of hormone-independent and hormone-responsive variants of MCF-7 human breast cancer cells. Clin. Exp. Metastasis, *11:* 15–26, 1993.
43. Rochefort, H., Liaudet, E., and Garcia, M. Alterations and role of human cathepsin D in cancer metastasis. Enzyme Protein, *49:* 106–116, 1996.
44. Keeling, J., and McKee, G. T. Heat shock protein (HSP)27: a further refinement in the diagnosis of suspicious fine needle aspirates of breast. Cytopathology, *10:* 40–49, 1999.
45. Ogawa, K., Kudo, H., Kim, Y. C., Nakashima, Y., Ohshio, G., and Yamabe, H. Expression of vitamin B12 R-binder in breast tumors. An immunohistochemical study. Arch. Pathol. Lab. Med., *112:* 1117–1120, 1988.
46. Wu, K., Helzlsouer, K. J., Comstock, G. W., Hoffman, S. C., Nadeau, M. R., and Selhub, J. A prospective study on folate, B12, and pyridoxal 5′-phosphate (B6) and breast cancer. Cancer Epidemiol. Biomark. Prev., *8:* 209–217, 1999.
47. Clark, G. J., and Der, C. J. Aberrant function of the Ras signal transduction pathway in human breast cancer. Breast Cancer Res. Treat., *35:* 133–144, 1995.
48. Guner, G., Kirkali, G., Yenisey, C., and Tore, I. R. Cytosol and serum ferritin in breast carcinoma. Cancer Lett., *67:* 103–112, 1992.
49. Kerr, M. K., and Churchill, G. A. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc. Natl. Acad. Sci. USA, *98:* 8961–8965, 2001.
50. Novak, J. P., Sladek, R., and Hudson, T. J. Characterization of variability in large-scale gene expression data: implications for study design. Genomics, *79:* 104–113, 2002.
51. Dotzlaw, H., Miller, T., Karvelas, J., and Murphy, L. C. Epidermal growth factor gene expression in human breast cancer biopsy samples: relationship to estrogen and progesterone receptor gene expression. Cancer Res., *50:* 4204–4208, 1990.
52. Tsai, J. C., Liu, L., Guan, J., and Aird, W. C. The egr-1 gene is induced by epidermal growth factor in ECV304 cells and primary endothelial cells. Am. J. Physiol. Cell Physiol., *279:* C1414–C1424, 2000.
53. Das, A., Chendil, D., Dey, S., Mohiuddin, M., Mohiuddin, M., Milbrandt, J. D., Rangnekar, V. M., and Ahmed, M. M. Ionizing radiation down-regulates p53 protein in primary Egr-1−/− mouse embryonic fibroblast cells causing enhanced resistance to apoptosis. J. Biol. Chem., 2000.
54. Chapman, N. R., and Perkins, N. D. Inhibition of the RelA(p65) NF-κB subunit by Egr-1. J. Biol. Chem., *275:* 4719–4725, 2000.
55. Huang, R. P., Fan, Y., de Belle, I., Niemeyer, C., Gottardis, M. M., Mercola, D., and Adamson, E. D. Decreased Egr-1 expression in human, mouse and rat mammary cells and tissues correlates with tumor formation. Int. J. Cancer, *72:* 102–109, 1997.
56. Tanaka, N., Ishihara, M., Kitagawa, M., Harada, H., Kimura, T., Matsuyama, T., Lamphier, M. S., Aizawa, S., Mak, T. W., and Taniguchi, T. Cellular commitment to oncogene-induced transformation or apoptosis is dependent on the transcription factor IRF-1. Cell, *77:* 829–839, 1994.
57. Tanaka, N., Ishihara, M., Lamphier, M. S., Nozawa, H., Matsuyama, T., Mak, T. W., Aizawa, S., Tokino, T., Oren, M., and Taniguchi, T. Cooperation of the tumour suppressors IRF-1 and p53 in response to DNA damage. Nature (Lond.), *382:* 816–818, 1996.
58. Mori, K., Stone, S., Khaodhiar, L., Braverman, L. E., and DeVito, W. J. Induction of transcription factor interferon regulatory factor-1 by interferon-γ (IFN γ) and tumor necrosis factor-α (TNF α) in FRTL-5 cells. J. Cell Biochem., *74:* 211–219, 1999.
59. Burow, M. E., Weldon, C. B., Tang, Y., Navar, G. L., Krajewski, S., Reed, J. C., Hammond, T. G., Clejan, S., and Beckman, B. S. Differences in susceptibility to tumor necrosis factor α-induced apoptosis among MCF-7 breast cancer cell variants. Cancer Res., *58:* 4940–4946, 1998.

60. Egeblad, M., and Jaattela, M. Cell death induced by TNF or serum starvation is independent of ErbB receptor signaling in MCF-7 breast carcinoma cells. Int. J. Cancer, *86:* 617–625, 2000.
61. Sheen-Chen, S. M., Chen, W. J., Eng, H. L., and Chou, F. F. Serum concentration of tumor necrosis factor in patients with breast cancer. Breast Cancer Res. Treat., *43:* 211–215, 1997.
62. Zhou, B. P., Hu, M. C., Miller, S. A., Yu, Z., Xia, W., Lin, S. Y., and Hung, M. C. HER-2/neu blocks tumor necrosis factor-induced apoptosis via the Akt/NF-κB pathway. J. Biol. Chem., *275:* 8027–8031, 2000.
63. Zyad, A., Bernard, J., Clarke, R., Tursz, T., Brockhaus, M., and Chouaib, S. Human breast cancer cross-resistance to TNF and adriamycin: relationship to MDR1. Mn-SOD and TNF gene expression. Cancer Res., *54:* 825–831, 1994.
64. Schiff, R., Reddy, P., Ahotupa, M., Coronado-Heinsohn, E., Grim, M., Hilsenbeck, S. G., Lawrence, R., Deneke, S., Herrera, R., Chamness, G. C., Fuqua, S. A., Brown, P. H., and Osborne, C. K. Oxidative stress and AP-1 activity in tamoxifen-resistant breast tumors *in vivo.* J. Natl. Cancer Inst., *92:* 1926–1934, 2000.
65. Nakshatri, H., Bhat-Nakshatri, P., Martin, D. A., Goulet, R. J., and Sledge, G. W. Constitutive activation of NF-κB during progression of breast cancer to hormone-independent growth. Mol. Cell. Biol., *17:* 3629–3639, 1997.
66. Clarkson, R. W., and Watson, C. J. NF-κB and apoptosis in mammary epithelial cells. J. Mammary Gland Biol. Neoplasia, *4:* 165–175, 1999.
67. van der Burg, B., Slager-Davidov, R., van der Leede, B. M., de Laat, S. W., and ven der Saag, P. T. Differential regulation of AP1 activity by retinoic acid in hormone-dependent and -independent breast cancer cells. Mol. Cell. Endocrinol., *112:* 143–152, 1995.
68. Hehner, S. P., Hofmann, T. G., Droge, W., and Schmitz, M. L. The antiinflammatory sesquiterpene lactone parthenolide inhibits NF-κB by targeting the IκB kinase complex. J. Immunol., *163:* 5617–5623, 1999.
69. Patel, N. M., Nozaki, S., Shortle, N. H., Bhat-Nakshatri, P., Newton, T. R., Rice, S., Gelfanov, V., Boswell, S. H., Goulet, R. J., Jr., Sledge, G. W., Jr., and Nakshatri, H. Paclitaxel sensitivity of breast cancer cells with constitutively active NF-κB is enhanced by IκBα super-repressor and parthenolide. Oncogene, *19:* 4159–4169, 2000.
70. Garcia-Pineres, A. J., Castro, V., Mora, G., Schmidt, T. J., Strunck, E., Pahl, H. L., and Merfort, I. Cysteine 38 in p65/NF-κB plays a crucial role in DNA binding inhibition by sesquiterpene lactones. J. Biol. Chem., *276:* 39713–39720, 2001.
71. Taniguchi, T. Transcription factors IRF-1 and IRF-2: Linking the immune responses and tumor suppression. J. Cell Physiol., *173:* 128–130, 1997.
72. Tamura, T., Ishihara, M., Lamphier, M. S., Tanaka, N., Oishi, I., Aizawa, S., Matsuyama, T., Mak, T. W., Taki, S., and Taniguchi, T. An IRF-1-dependent pathway of DNA damage-induced apoptosis in mitogen-activated T lymphocytes. Nature (Lond.), *376:* 596–599, 1995.
73. Sanceau, J., Hiscott, J., Delattre, O., and Wietzerbin, J. IFN-β induces serine phosphorylation of Stat-1 in Ewing's sarcoma cells and mediates apoptosis via induction of IRF-1 and activation of caspase-7. Oncogene, *19:* 3372–3383, 2000.
74. Boudreau, N., Sympson, C. J., Werb, Z., and Bissell, M. J. Suppression of ICE and apoptosis in mammary epithelial cells by extracellular matrix. Science (Wash. DC), *267:* 891–893, 1995.
75. Keane, M. M., Ettenberg, S. A., Lowrey, G. A., Russell, E. K., and Lipkowitz, S. Fas expression and function in normal and malignant breast cell lines. Cancer Res., *56:* 4791–4798, 1996.
76. Skaar, T. C., Bouker, K. B., and Clarke, R. Interferon regulatory factor-1 (IRF-1) in breast cancer. Proc. Am. Assoc. Cancer Res. *41:* 428, 2000.
77. Kondo, T., Minamino, N., Nagamura-Inoue, T., Matsumoto, M., Taniguchi, T., and Tanaka, N. Identification and characterization of nucleophosmin/B23/numatrin which binds the anti-oncogenic transcription factor IRF-1 and manifests oncogenic activity. Oncogene, *15:* 1275–1281, 1997.
78. Wang, C. Y., Cusack, J. C., Jr., Liu, R., and Baldwin, A. S., Jr. Control of inducible chemoresistance: enhanced anti-tumor therapy through increased apoptosis by inhibition of NF-κB. Nat. Med., *5:* 412–417, 1999.
79. Kim, D. W., Sovak, M. A., Zanieski, G., Nonet, G., Romieu-Mourez, R., Lau, A. W., Hafer, L. J., Yaswen, P., Stampfer, M., Rogers, A. E., Russo, J., and Sonenshein, G. E. Activation of NF-κB/Rel occurs early during neoplastic transformation of mammary cells. Carcinogenesis (Lond.). *21:* 871–879, 2000.
80. Welsh, P. L., and King, M. C. BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. Hum. Mol. Genet., *10:* 705–713, 2001.
81. Blum, D., Torch, S., Nissou, M. F., and Verna, J. M. 6-hydroxydopamine-induced nuclear factor-κ B activation in PC12 cells. Biochem. Pharmacol., *62:* 473–481, 2001.
82. Cavallini, L., Francesconi, M. A., Zoccarato, F., and Alexandre, A. Involvement of nuclear factor-κ B (NF-κB) activation in mitogen-induced lymphocyte proliferation: inhibitory effects of lymphoproliferation by salicylates acting as NF-κB inhibitors. Biochem. Pharmacol., *62:* 141–147, 2001.
83. Habener, J. F. Cyclic AMP response element binding proteins: a cornucopia of transcription factors. Mol. Endocrinol., *4:* 1087–1094, 1990.
84. Borrelli, E., Montmayeur, J. P., Foulkes, N. S., and Sassone-Corsi, P. Signal transduction and gene control: the cAMP pathway. Crit. Rev. Oncog., *3:* 321–338, 1992.
85. Cho, H., Aronica, S. M., and Katzenellenbogen, B. S. Regulation of progesterone receptor gene expression in MCF-7 breast cancer cells: a comparison of the effects of cyclic adenosine 3′, 5′- monophosphate, estradiol, insulin-like growth factor-I, and serum factors. Endocrinology, *134:* 658–664, 1994.
86. Lee, Y. N., Park, Y. G., Choi, Y. H., Cho, Y. S., and Cho-Chung, Y. S. CRE-transcription factor decoy oligonucleotide inhibition of MCF-7 breast cancer cells: cross-talk with p53 signaling pathway. Biochemistry, *39:* 4863–4868, 2000.
87. Drew, P. D., Franzoso, G., Becker, K. G., Bours, V., Carlson, L. M., Siebenlist, U., and Ozato, K. NF κ B and interferon regulatory factor 1 physically interact and synergistically induce major histocompatibility class I gene expression. J. Interferon Cytokine Res., *15:* 1037–1045, 1995.
88. Neish, A. S., Read, M. A., Thanos, D., Pine, R., Maniatis, T., and Collins, T. Endothelial interferon regulatory factor 1 cooperates with NF-κ B as a transcriptional activator of vascular cell adhesion molecule 1. Mol. Cell. Biol., *15:* 2558–2569, 1995.
89. Coccia, E. M., Del Russo, N., Stellacci, E., Orsatti, R., Benedetti, E., Marziali, G., Hiscott, J., and Battistini, A. Activation and repression of the 2–5A synthetase and p21 gene promoters by IRF-1 and IRF-2. Oncogene, *18:* 2129–2137, 2000.
90. Escalante, C. R., Yie, J., Thanos, D., and Aggarwal, A. K. Structure of IRF-1 bound DNA reveals determinants of interferon regulation. Nature (Lond.), *391:* 103–106, 1998.
91. Lee, A. H., Hong, J-H., and Seo, Y. S. Tumor necrosis factor-α and interferon-γ synergistically activate the RANTES promoter through nuclear factor κB and interferon regulatory factor 1 (IRF-1) transcription factors. Biochem. J., *350:* 131–138, 2000.
92. Sanceau, J., Kaisho, T., Hirano, T., and Wietzerbin, J. Triggering of the human interleukin-6 gene by interferon-γ and tumor necrosis factor-α in monocytic cells involves cooperation between interferon regulatory factor-1, NFκB, and SP1 transcription factors. J. Biol. Chem., *270:* 27920–27931, 1995.
93. Kim, T. K., and Maniatis, T. The mechanism of transcriptional synergy of an *in vitro* assembled interferon-β enhanceosome. Mol. Cell, *1:* 119–129, 1997.
94. van den Berg, H. W., Leahey, W. J., Lynch, M., Clarke, R., and Nelson, J. Recombinant human interferon α increases oestrogen receptor expression in human breast cancer cells (ZR-75–1) and sensitises them to the anti-proliferative effects of tamoxifen. Br. J. Cancer, *55:* 255–257, 1987.
95. Lindner, D. J., and Borden, E. C. Effects of tamoxifen and interferon-β or the combination on tumor-induced angiogenesis. Int. J. Cancer, *71:* 456–461, 1997.
96. Buzzi, E., Brugia, M., Trippa, F., Rossi, G., Trivisonne, R., Giustini, L., Pinaglia, D., Capparella, V., and Sica, G. Natural interferon-β and tamoxifen in hormone-resistant patients with advanced breast cancer. Anticancer Res., *15:* 2187–2190, 1995.
97. Delgado, M., Munoz-Elias, E. J., Gomariz, R. P., and Ganea, D. Vasointestinal polypeptide and pituitary adenylate cyclase-activating polypeptide prevent inducible nitric oxide synthase transcription in macrophages by inhibiting NFκB and IFN regulatory factor 1 activation. J. Immunol., *162:* 4685–4696, 1999.
98. Clarke, R., Dickson, R. B., and Lippman, M. E. Hormonal aspects of breast cancer: growth factors, drugs and stromal interactions. Crit. Rev. Oncol. Hematol., *12:* 1–23, 1992.

Development and Validation of a Method for Using Breast Core Needle Biopsies for Gene Expression Microarray Analyses[1]

Matthew Ellis,[2] Natalie Davis, Andrew Coop, Minetta Liu, Lisa Schumaker, Richard Y. Lee, Rujirutana Srikanchana, Chris G. Russell, Baljit Singh, William R. Miller, Vered Stearns,[3] Marie Pennanen, Theodore Tsangaris, Ann Gallagher, Aiyi Liu, Alan Zwart, Daniel F. Hayes,[3] Marc E. Lippman,[3] Yue Wang, and Robert Clarke[4]

Departments of Oncology [M. E., N. D., A. C., M. L., L. S., R. Y. L., V. S., A. G., A. Z., D. F. H., M. E. L., R. C.], Pathology [B. S.], Surgery [M. P., T. T.], Physiology and Biophysics [R. Y. L., R. C.], and Biostatistics and Biomathematics [A. L.], Georgetown University School of Medicine, Washington, DC 20007; Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064 [R. S., Y. W.]; ResGen, part of the Invitrogen Corporation, Huntsville, Alabama 35801 [C. G. R.]; and Department of Oncology, University of Edinburgh, Western General Hospital, Edinburgh, Scotland, EH4 2XU United Kingdom [W. R. M.]

ABSTRACT

Purpose: **Gene expression microarray technologies have the potential to define molecular profiles that may identify specific phenotypes (diagnosis), establish a patient's expected clinical outcome (prognosis), and indicate the likelihood of a beneficial effect of a specific therapy (prediction). We wished to develop optimal tissue acquisition, processing, and analysis procedures for exploring the gene expression profiles of breast core needle biopsies representing cancer and noncancer tissues.**

Experimental Design: **Human breast cancer xenografts were used to evaluate several processing methods for prospectively collecting adequate amounts of high-quality RNA for gene expression microarray studies. Samples were assessed for the preservation of tissue architecture and the quality and quantity of RNA recovered. An optimized protocol was applied to a small study of core needle breast biopsies from patients, in which we compared the molecular profiles from cancer with those from noncancer biopsies. Gene expression data were obtained using Research Genetics, Inc. NamedGenes cDNA microarrays. Data were visualized using simple hierarchical clustering and a novel principal component analysis-based multidimensional scaling. Data dimensionality was reduced by simple statistical approaches. Predictive neural networks were built using a multilayer perceptron and evaluated in an independent data set from snap-frozen mastectomy specimens.**

Results: **Processing tissue through RNA*Later* preserves tissue architecture when biopsies are washed for 5 min on ice with ice-cold PBS before histopathological analysis. Cell margins are clear, tissue folding and fragmentation are not observed, and integrity of the cores is maintained, allowing optimal pathological interpretation and preservation of important diagnostic information. Adequate concentrations of high-quality RNA are recovered; 51 of 55 biopsies produced a median of 1.34 μg of total RNA (range, 100 ng to 12.60 μg). Snap-freezing or the use of RNA*Later* does not affect RNA recovery or the molecular profiles obtained from biopsies. The neural network predictors accurately discriminate between predominantly cancer and noncancer breast biopsies.**

Conclusions: **The approaches generated in these studies provide a simple, safe, and effective method for prospectively acquiring and processing breast core needle biopsies for gene expression studies. Gene expression data from these studies can be used to build accurate predictive models that separate different molecular profiles. The data establish the use and effectiveness of these approaches for future prospective studies.**

Received 11/27/01; revised 2/4/02; accepted 2/15/02.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] Supported by USPHS from National Cancer Institute Grants 5R33CA83231 (to Y. W.), R01-CA/AG58022 (to R. C.), P50-CA58185 (to R. C.), and K12-CA76903 (to M. L.), the Department of Defense Grants BC980629 (to R. C.) and BC990358 (to R. C.), and CI-3 Cancer Research Fund-Lilly Clinical Investigator Award of the Damon Runyon-Walter Winchell Foundation (to V. S., D. F. H.).

[2] Present address: Breast Cancer Program, 25105, The Morris Building, Duke University Medical Center, Durham, NC 27710.

[3] Present address: Department of Medicine, 3110 Taubman Centre, University of Michigan, Ann Arbor, MI 48109.

[4] To whom requests for reprints should be addressed, at Department of Oncology, Georgetown University School of Medicine, W405A Research Building, 3970 Reservoir Road NW, Washington, DC 20007. Phone: (202) 687-3755; Fax: (202) 687-7505; E-mail: clarker@georgetown.edu.

INTRODUCTION

The emerging gene microarray technologies provide powerful new methodologies with which to address several important issues in breast cancer research. For example, it should be possible to define gene expression patterns that can identify specific phenotypes (diagnosis), establish a patient's expected clinical outcome (prognosis), and indicate the likelihood of a beneficial effect of a specific therapy (prediction; Refs. 1 and 2). Gene microarray technologies are performed on chips, glass slides, or filters and allow the comparison of gene expression profiles from two or more tissues or the same tissue in different biological states (3). The technologies continue to develop, with considerable discussion regarding which technology has the greatest potential to address the molecular profiling of tumors. Each of the major approaches has advantages and disadvan-

tages, but the most important consideration is the ability of the technology to address the chosen hypothesis (4). Overall, there is no compelling evidence of major differences in the accuracy or reproducibility of the various microarray platforms (4–6). Studies that directly compare the nylon-based cDNA arrays with either glass slide cDNA arrays and/or oligonucleotide chips consistently report that these platforms produce comparable data (5–8).

Because gene expression technologies provide an assessment of mRNA abundance in a sample, all require the production of a probe, labeled with either a radioactive nucleotide or fluorescent molecule, generated from either the total or polyadenylate RNA isolated from the sample. Currently, it is not possible to isolate adequate concentrations of high-quality RNA from what would otherwise be the most abundant source: the formalin-fixed, paraffin-embedded tumor specimens available in established tumor banks. Only fresh or appropriately frozen tissues provide the necessary quality of RNA for the preparation of probes to hybridize to existing gene expression microarrays.

Whereas many institutions have frozen tumor banks, these may be of limited use in obtaining reproducible gene expression profiles for some breast cancers. For example, most are heavily biased toward large breast tumors (T_3–T_4). These tumors are poorly representative of the small tumors now seen in many patients for initial diagnosis (9). A further concern with existing frozen tissue banks is the frequent lack of a standardized approach for tissue acquisition and processing. Tissue handling between excision and freezing can vary considerably. For example, some tumors are frozen within seconds of excision, and others are placed on wet or dry ice after excision, whereas some may stand for many minutes at room temperature before being placed in liquid nitrogen. The importance of tissue processing is often critical for assessing various end points and can affect both RNA stability for RNA *in situ* hybridizations and antigen stability/accessibility for immunohistochemistry (10).

The effect of tissue acquisition and processing on gene microarray data has not been widely addressed. Nonetheless, this is likely to be important for at least two critical parameters. First is preservation of high-quality RNA. Most investigators acknowledge the importance of using only pure, high-quality RNA for gene microarray studies (11). The second factor is maintenance of a tissue's gene expression profile. For example, hypoxia- or stress-induced responses can be induced in metabolically active cells. Oxygen deprivation begins with the loss of tissue perfusion occurring upon excision. This deprivation can trigger a hypoxic response, characterized by the altered expression of specific genes (12, 13). Several of these genes are transcription factors that further affect the expression of their target genes (13).

One problem with these two factors is that both can affect a sample, but RNA could still be obtained, a probe could still be generated, and a molecular profile could still be obtained after hybridization to a gene expression microarray. Subtle changes that are time, temperature, pH, and/or oxygen dependent could occur with sufficient variability that they are almost impossible to detect reproducibly. Some tumors with high metabolic activity may be more sensitive to hypoxia, producing a statistically valid and biologically plausible clustering that could have resulted more from tissue processing rather than tissue biology. Where such changes are subtle, expression profiles might still appear grossly similar, complicating an assessment of tissue processing artifacts.

Table 1 Experimental conditions for xenograft study
All samples were processed in duplicate.

Temperature of 72 h storage	Wash solution	Wash time
4°C	1:6 (RNA*Later*:PBS)	5 min
4°C	1:9 (RNA*Later*:PBS)	5 min
4°C	1:12 (RNA*Later*:PBS)	5 min
4°C	PBS	5 min
4°C	No wash	No wash
4°C	1:6 (RNA*Later*:PBS)	120 min
4°C	1:9 (RNA*Later*:PBS)	120 min
4°C	1:12 (RNA*Later*:PBS)	120 min
4°C	PBS	120 min
Room temperature	No wash	No wash
−20°C	No wash	No wash

Given the bias of existing banks and the potential differences in tissue processing, many important questions in breast cancer biology may require prospective study designs. Such study designs are more valid for the exploration or validation of new predictive and prognostic factors. Whereas optimized tissue acquisition and processing strategies for prospective studies offer the opportunity for greater control of tissue quality than retrospective studies, these strategies have not been described. In this study, we wished to develop a standard tissue acquisition/processing method for prospective core needle breast biopsy sampling. This method should avoid the initial use of liquid nitrogen, preserve tissue architecture, and provide adequate concentrations of high-quality RNA for microarray analysis. We now report a simple tissue processing approach using a commercially available reagent (RNA*Later*) that is applicable to prospective studies on core needle biopsies. RNA obtained from this approach was compared with RNA from snap-frozen human breast biopsies of neoplastic and nonneoplastic tissues, gene expression microarray data were obtained, and an accurate neural network capable of discriminating between these tissues was built and validated in an independent data set.

MATERIALS AND METHODS

Breast Cancer Xenograft Studies. MDA-MB-231 cells were inoculated into athymic nude mice as described previously (14, 15). Mice were sacrificed, and tumor tissue was obtained using sterile scissors and forceps. Needle biopsies were taken from the excised xenografts and placed into separate tubes containing 0.5 ml of RNA*Later* (Ambion, Austin, TX) at room temperature. Samples were stored at various temperatures for 72 h and subsequently processed according to the scheme in Table 1. Each experimental condition was explored in duplicate samples. Tissues were embedded in OCT (BDH; Poole, Dorset, United Kingdom), and standard frozen sections were prepared from each sample. Subsequently, sections were stained with H&E and evaluated by the study pathologist. The remainder of the core was stored at −80°C, and total RNA was extracted for evaluation. All animal studies were performed under protocols approved

by the Georgetown University Animal Care and Use Committee.

Patient Population. Patients undergoing a diagnostic core needle or excisional biopsy at Georgetown University Hospital were eligible for the tissue acquisition protocol, in which additional cores were obtained for study purposes. All patients signed a written consent form approved by the Georgetown University Medical Center Institutional Review Board. Core biopsies provided by the radiologists were obtained with either mammographic or ultrasound guidance. Core biopsies obtained by the surgeons were obtained either after surgical exposure of the tumor or during a routine needle biopsy. A total of 1–4 cores were obtained from each patient for study purposes, depending on the size of the breast lesion. In addition, nine frozen breast tumor specimens were obtained from the Department of Oncology, University of Edinburgh (Edinburgh, Scotland, United Kingdom) for use in testing the neural networks for accuracy in identifying tissues as malignant or nonmalignant. These samples were collected after appropriate patient consent and consistent with the relevant United Kingdom legislation. In this study, the pathologist was blinded to all clinical information on all samples.

Collection and Handling of Human Breast Core Biopsies for Microarray Analysis. Generally, 1–4 core needle biopsies (14-gauge needle) were obtained from each consenting patient. Random cores were immediately snap-frozen in liquid nitrogen; others were individually placed in separate cryo-tubes containing 0.5 ml of RNA*Later* solution. Snap-frozen tissues were placed directly in liquid nitrogen from the core biopsy needle, immediately upon removal from the patient. For the RNA*Later* samples, core biopsies were placed in 500 μl of RNA*Later* and maintained at 4°C for 24 h before snap-freezing. Each tube was labeled with the patient's name, hospital number, and study number. Frozen samples were transferred to the Lombardi Cancer Center's Tissue and Histopathology Shared Resource (Washington, DC) for processing.

Before removing the samples from the tube for frozen section preparation, each sample was washed for 5 min on ice with 500 μl of ice-cold sterile PBS (RNase free); otherwise, samples in RNA*Later* will not freeze in the cryostat. Each core biopsy sample was then embedded separately in an OCT block. A frozen section was taken, stained with H&E, and examined by the study pathologist. OCT-embedded samples were maintained frozen at −70°C until the analysis of the main tumor mass was complete.

The study pathologist evaluated all biopsies to determine the presence of invasive cancer and to estimate the relative amounts of normal epithelium, stroma, and fat. Because samples were to be used for microarray analysis, the percentage of invasive cancer, normal epithelium, stroma, and fat was estimated relative to cell nuclei only. Provided this histological review offered no new clinical information important for patient care, biopsies suitable for microarray were identified. In this manner, tissue for expression microarray analysis was ensured to be of no new diagnostic relevance. This determination is important because RNA extraction destroys tissue architecture. If the samples had contained information that modified the surgical pathology diagnosis, these biopsies would not have been used. This situation did not occur in this study.

Once released for study, all patient identifiers were removed from each sample. The link between patient identifiers and study identifiers was held in a confidential database. Access to this database was reserved only for the clinical study principal investigator and the data entry technician. The frozen clinical material, mostly frozen in OCT, was directly provided to the research laboratory for storage and/or processing. Upon receipt in the research laboratory, tissue was either stored at −80°C or processed immediately for RNA extraction.

Preparation and Quality Assessment of RNA from Frozen Tissues. Frozen tissue was placed in a 1 × 1-inch plastic bag on dry ice and pulverized, and lysis buffer from the Qiagen RNeasy kit was added (Qiagen, Inc., Valencia, CA). Each sample was then transferred to a 1.5-ml centrifuge tube, homogenized with a 1-ml syringe and an 18-gauge needle, added to the Qiagen spin column, and centrifuged to bind the RNA to the matrix. The column was washed with the buffers provided in the kit, and the RNA was finally eluted with distilled H_2O. RNA concentrations were determined by comparing the absorbance ratios ($A_{260\ nm}/A_{280\ nm}$) obtained spectrophotometrically using a Beckman DU640 Spectrophotometer (Beckman, Fullerton, CA).

Because using standard gel electrophoresis to assess RNA quality would require almost the entire RNA sample, we used an Agilent 2100 analyzer and RNA 6000 LabChip kits (RNA microelectroseparation and analysis; Agilent Technologies, New Castle, DE). A total of 100 ng of each RNA sample was loaded/well. The analyzer allows for visual examination of both the 18S and 28S rRNA bands as a measure of RNA integrity.

Probe Generation for Gene Microarray Hybridizations. Probes were generated as described previously (16). This method radiolabels both the sense and antisense probe strands and further increases probe-specific activity by incorporating two radiolabeled nucleotides. Thus, tumors can be arrayed on nylon filter arrays with as little as 100 ng of total RNA and without RNA amplification (7, 16). Whereas an adequate signal is generated with 100 ng of total RNA, the use of very low RNA concentrations will likely affect the ability to adequately and reproducibly detect many lower abundance mRNAs. We used 500 ng of total RNA, which is sufficient to allow the use of approximately 70% of breast needle biopsies without either RNA amplification or pooling. None of the RNAs was amplified or pooled in the current study.

To synthesize the labeled cDNA probe, 500 ng of total RNA were incubated at 70°C for 10 min with 2 mg of oligodeoxythymidylate and then chilled on ice for 2 min. The primed DNA was incubated at 37°C for 90 min in a solution containing 1× first strand, 3 mM DTT, 1 mM dGTP/dTTP, 300 units of reverse transcriptase, 50 mCi of [^{33}P]dCTP, and 50 mCi of [^{33}P]dATP. The second strand was synthesized by adding 1× reaction buffer, 100 units of DNA polymerase I, 500 ng of random primers, 1 mM dGTP/dTTP, 50 mCi of [^{33}P]dCTP, and 50 mCi of [^{33}P]dATP. The reaction was incubated for 2 h at 16°C. A radiolabeled probe was purified using a BioSpin-6 chromatography column (Bio-Rad) and denatured by boiling for 3 min. A purified probe was added to the hybridization roller tube containing the prehybridized GeneFilter and incubated for 12–18 h at 42°C in a Robin Scientific Roller Oven. For these studies, the NamedGenes

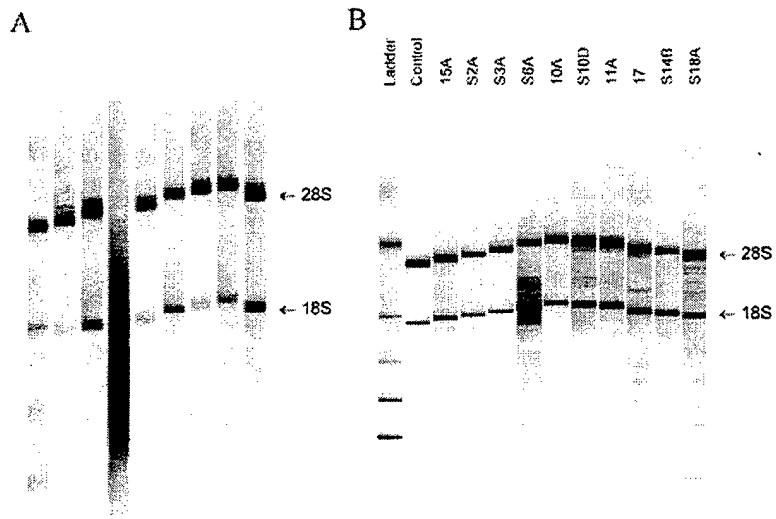


Fig. 1 The quality of RNA recovered from human core breast needle biopsies. RNA was evaluated using an Agilent 2100 analyzer. *A*, xenografts; *B*, breast core needle biopsies arrayed in Figs. 5 and 6 and further characterized in Table 4. In the images displayed, fluorescence scales are not equivalent between lanes but have been normalized for clarity.

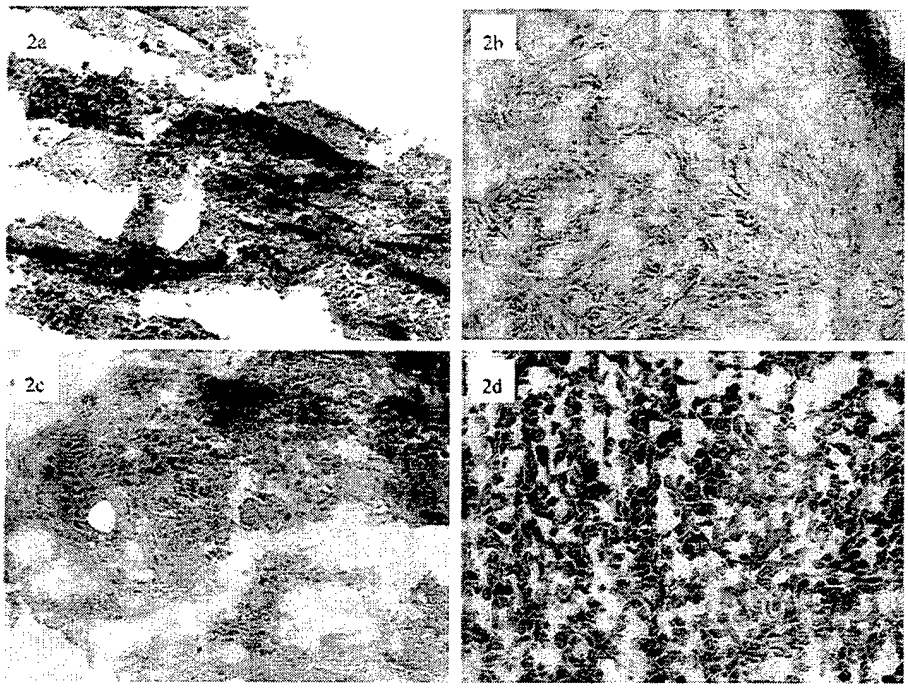


Fig. 2 MDA-MB-231 human breast tumor xenografts processed for frozen tissue sectioning in RNA*Later*. *A*, no wash; *B*, washed in PBS:RNA*Later* (1:6; v/v) for 2 h at 4°C; *C*, washed in PBS:RNA*Later* (1:6; v/v) for 5 min at 4°C; *D*, washed in PBS for 5 min on ice.

filters (Research Genetics, Inc., Huntsville, AL) were used. These filters contain 4032 known genes, 192 housekeeping genes, and 192 control genes on each filter. Each hybridized GeneFilter was washed twice in 2× SSC, 1% SDS at 50°C for 20 min and once at 55°C in 0.5× SSC, 1% SDS for 15 min. Hybridization signals were detected by phosphorimaging using a Molecular Dynamics Storm PhosphorImager (Molecular Dynamics, Sunnyvale, CA). The sensitivity and reproducibility of these and other nylon filter-based cDNA microarrays have been widely reported (7, 17–20).

Normalization of Data. Pathways software algorithms (Research Genetics, Inc.) were used to correct for nonspecific binding of the probe to filter (background correction). Approaches for signal normalization, intended to correct for differences in probe specific activities, hybridizations, and other interexperiment variables, are diverse (11). In the present study, the average of all data points was used to calculate a normalization factor; the normalized intensity value for each spot was obtained by multiplying the normalization factor by the raw intensity (11).

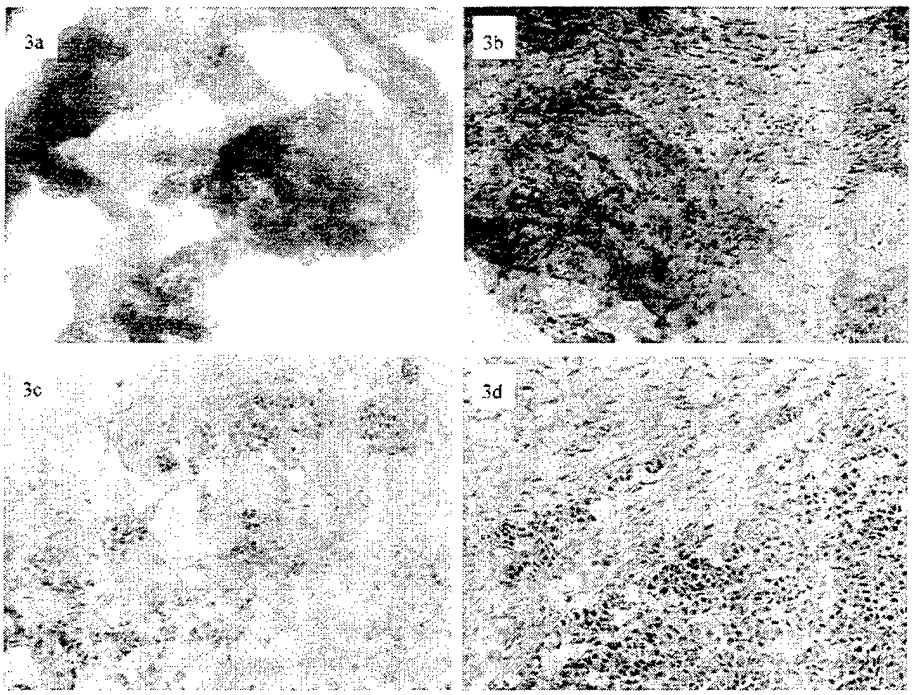


Fig. 3 Human breast needle biopsies processed for frozen tissue sectioning in RNA*Later*. *A*, no wash; *B*, washed in PBS: RNA*Later* (1:6; v/v) for 2 h at 4°C; *C*, washed in PBS:RNA-*Later* (1:6; v/v) for 5 min at 4°C; *D*, washed in PBS for 5 min on ice.

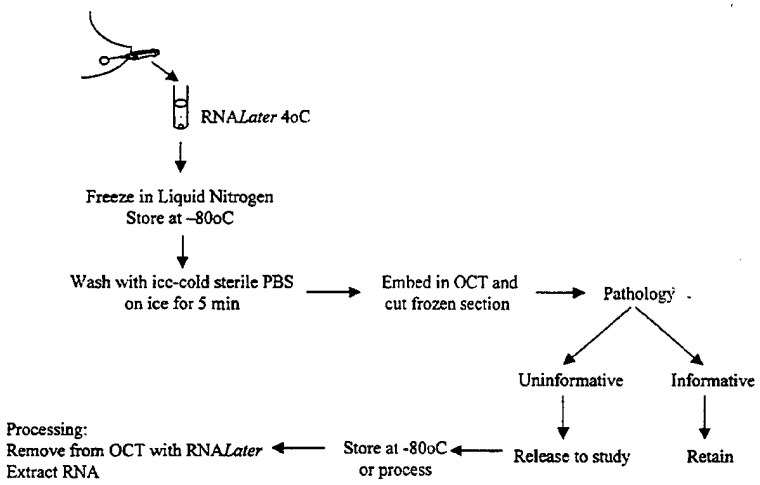


Fig. 4 Optimized tissue acquisition/processing procedure for breast needle biopsies.

Analysis of Gene Microarray Data. The optimal approach for analyzing the high dimensional gene expression data generated by gene microarray studies remains unclear. The high dimensionality of these data are problematic, with most existing analyses functioning more accurately in low dimensionality (21). However, rather than making statistical inference for identifying and studying functionally relevant genes, the study goal was to validate the tissue acquisition and processing methods and demonstrate the applicability of this approach for building clinically relevant predictive models.

Recently, we devised a simple approach to the exploration of small studies with two experimental groups.[5] Our approach used simple statistical analyses to reduce data dimensionality and identify subsets of discriminant genes. This approach is similar in principle to that used by Hedenfalk *et al.* (22). Because the class of each sample (cancer *versus* noncancer) is

[5] Z. Gu. Association of interferon regulatory factor-1, nucleophosmin, nuclear factor-kappa-B and cAMP binding with acquired resistance to Faslodex (ICI 182,780), submitted for publication.

Table 2 Concentration of RNA recovered from breast needle biopsies

Biopsies	RNA ≥ 100 ng[a]	Total RNA recovered ($\bar{X}$ ± SE)	Range (>100 ng)
$n = 55$[b]	51/55 (93%)	3.63 ± 0.48 μg; median = 1.34 μg	100 ng to 12.60 μg
$n = 25$	Snap-frozen	2.04 ± 0.51 μg; median = 1.32 μg	100 ng to 9.00 μg[c]
$n = 21$	RNA*Later*	3.49 ± 0.78 μg; median = 2.70 μg	100 ng to 12.60 μg

[a] Number producing ≥100 ng of RNA, the minimum useful concentration of RNA without amplification, irrespective of the tissue acquisition and processing method applied.
[b] We did not have complete data on processing for the first 9 of the 55 samples.
[c] $P = 0.13$; Mann-Whitney rank-sum test; RNA*Later versus* snap-frozen tissue.

Table 3 Characteristics of breast needle biopsy material

A. Biopsy	Source	ER/PR[a]	% Cancer	% Normal	% Fat	% CT	RNA (μg)[b]
15A[c]	Radiology	ND	0%	70%	0%	30%	3.28
S6A[c]	Surgery	ND	0%	0%	100%	0%	2.49
10A	Radiology	ND	0%	5%	45%	50%	2.07
11A	Radiology	ND	0%	20%	40%	40%	1.36
17[c]	Radiology	ND	2%	0%	90%	8%	6.70
S2A	Surgery	+/+	90%	0%	5%	5%	3.20
S3A[c]	Surgery	+/+	90%	0%	0%	10%	2.70
S10D[c]	Surgery	+/+	80%	0%	0%	20%	6.50
S14B[c]	Surgery	−/−	80%	0%	0%	20%	4.20
S18A	Surgery	+/−	90%	0%	0%	10%	1.70

B. RNA recovered from biopsies used in this study ($\bar{X}$ ± SE)[d]

No RNA*Later*	2.21 ± 0.52 μg total RNA[d]
RNA*Later*	3.83 ± 0.73 μg total RNA
Overall RNA recovered	3.10 ± 1.60 μg total RNA

C. Case	Biopsies	Pathological diagnosis
S2	S2A	Invasive adenocarcinoma
	S2B	Invasive adenocarcinoma
	S2C	Invasive adenocarcinoma
S6	S6A	No cancer
	S6B	No cancer
	S6C	No tissue
S10	S10A	No cancer
	S10B	Possible DCIS
	S10C	No cancer
	S10D	Invasive adenocarcinoma

[a] PR, progesterone receptor; CT, connective tissue; DCIS, ductal carcinoma *in situ*; ND, not determined.
[b] Total RNA recovered from each needle biopsy. Five hundred ng of each RNA population were used to generate the probes hybridized to obtain the data presented in Fig. 5.
[c] Biopsies processed in RNA*Later*.
[d] $P = 0.129$; Student's t test; RNA*Later versus* no RNA*Later*.

known from the histopathological analyses, dimensionality can be reduced in a supervised manner by performing a series of statistical tests. The major purpose of performing these tests was only to select a group of genes that would be used for data visualization and analysis. Student's t test and a t test for unequal variances (each assumes normal distribution of the data) and a nonparametric (distribution-free) Wilcoxon test were used. Whereas the inflated type 1 error will overestimate significant differences, the incidence of false negative estimates should be smaller. Because the distribution of the data among and within replicate experiments and for individual genes cannot be determined (23), both logarithm-transformed and nontransformed data were compared.

Two reduced dimensional data sets were selected; one comprising genes with Ps < 0.05, and one comprising genes with Ps < 0.02. Because of their marked biological differences, these phenotypes should be easily separable. Thus, the data were visualized using our Fisher separability-based multidimensional scaling approach that projects high dimensional data into three-dimensional data space (24, 25). Because it has become widely used, visualization using the simple hierarchical clustering described by Eisen *et al.* (26) is also presented.

Generation and Testing of a Neural Network. To determine whether the genes we selected could be used to separate cancer from noncancer tissues, a neural network was trained using the gene expression microarray data from five cancer biopsies and five noncancer biopsies. Neural networks can be considered as parallel computing systems consisting of many simple processors with many interconnections. The main advantages of neural networks are that they can learn complex non-

linear input-output relationships, use sequential training procedures, and adapt themselves to the data (27, 28).

The learning process involves updating network architecture and connection weights so that the predictive model can efficiently perform a specific classification task. We used a multilayer preceptron to design a nonlinear neural classifier, using each of the gene's expression levels in the tissue samples as the input and the cancer *versus* noncancer phenotype of each sample as the output. Consequently, the network output comes to approximate the posterior Bayesian probabilities of a sample being either cancer or noncancer given its gene expression profile (27, 29, 30). Three experimental configurations were tested, with either the top 40, 80, and 103 dimensions (data set selecting the top 103 genes; $P < 0.05$) or 10, 20, and 30 dimensions (data set selecting the top 30 genes; $P < 0.02$). These top genes were selected based on their fold difference between cancer and noncancer and their respective *P*s. Two prediction models were built, one with 3 hidden nodes and 8 inputs and one with 5 hidden nodes and 18 inputs. Mean-squared error estimates were used to explore network performance. The "leave-one-out" method was used for the initial testing and training of each neural network (27, 29, 30).

RESULTS

RNA Quality and Tissue Architecture from Xenograft Tissues Processed Using RNA*Later*. Recovery of high-quality RNA was optimal when OCT-embedded tissue samples were removed from frozen blocks using a small volume of RNA*Later*, which thawed and softened the embedding medium before tissue extraction from the blocks. Thus, the frozen block was placed in a small plastic tray, with the embedded tissue facing up, and 500 μl of RNA*Later* were pipetted on top. Using this method, seven of eight samples yielded high-quality RNA (Fig. 1*A*; RNA integrity analyzed using the Agilent 2100 bioanalyzer and RNA 6000 LabChip kit). In previous experiments, where the OCT block was dissolved by vigorous shaking in a large volume of PBS and tissue fragments were recovered with a strainer, only 6 of 12 samples yielded fully intact RNA (data not shown).

Pathology was not interpretable from material frozen directly in RNA*Later* and transferred to OCT without wash steps (Fig. 2*A*). This reflects inadequate freezing in the cryostat and consequent tissue folding during the cutting process. When washed in PBS:RNA*Later* (1:6) for 2 h at 4°C, the tissue did not fold on cutting, but cell outlines appeared blurred, making pathological interpretation difficult (Fig. 2*B*). Washing in PBS:RNA*Later* (1:6) for 5 min at 4°C also eliminated tissue folding, but now the cell outlines appeared distinct. Nonetheless, tissue fragmentation occurred in some specimens, making pathological interpretation suboptimal (Fig. 2*C*). Optimal preservation of tissue architecture was obtained by washing tissue for 5 min on ice with ice-cold PBS. Cell margins were clear, tissue folding and fragmentation were not observed, and the integrity of the cores was maintained, allowing optimal pathological interpretation (Fig. 2*D*). Similar data were obtained from a human breast core biopsy released to this study (Fig. 3 *A–D*). A scheme of the optimized tissue acquisition protocol is shown in Fig. 4.

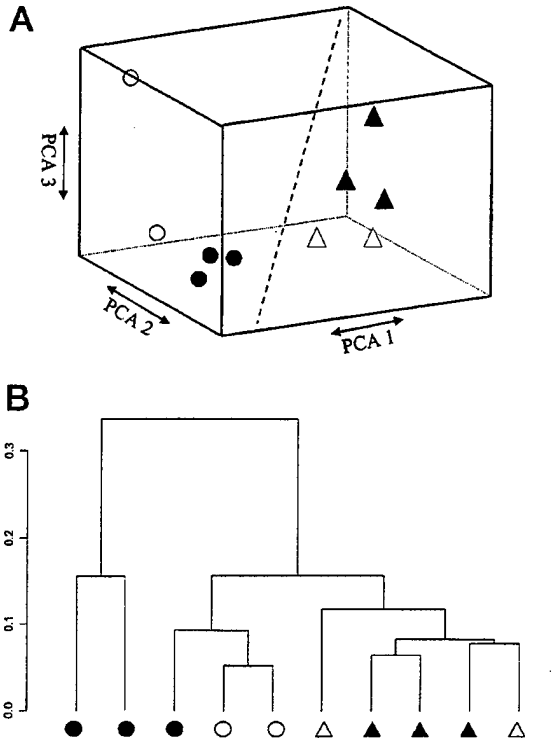


Fig. 5 Structure of the gene expression data using the top 103 genes. *A*, projection from top 103 genes (103 dimensions), selected by *t* test comparisons of $\log_{10}$-transformed gene expression data ($P \leq 0.05$) and projected into three dimensions; *B*, hierarchical clustering of samples in 2 dimensions based on 1-Pearson's coefficient matrix of the top 103 genes. Snap-frozen tissue: ○, noncancer; △, cancer. RNA*Later*-processed tissue: ●, noncancer; ▲, cancer.

Recovery of High-quality RNA from Human Breast Biopsies for Gene Microarray Studies. Cores were removed from OCT by placing the frozen block in a small plastic tray, with the embedded tissue facing up, and pipetting 500 μl of RNA*Later* on top. Intact cores were easily picked out of the OCT, which remained semisolid, using a sterile pipette tip. From a study of 55 breast needle biopsies, we obtained ≥100 ng of RNA on almost all samples (Table 2). The median value (1.34 μg) shows that most biopsies produce sufficient RNA to generate data using 500 ng of total RNA. There was no significant difference between frozen and RNA*Later*-processed biopsies in the mean concentrations of total RNA recovered (Tables 2 and 3). Thus, prospectively collected breast needle biopsies, either directly snap-frozen or processed in RNA*Later*, can produce adequate RNA concentrations for use in gene microarray studies.

A further requirement of gene expression microarray experiments is the isolation of high-quality RNA (11). Sufficient RNA was not recovered to allow for an assessment of RNA quality by both standard gel electrophoresis methods and gene microarray studies on the same samples. Because gel electrophoresis requires ~1 μg of RNA, the Agilent 2100 "lab-on-a-chip" technology was used to assess RNA quality. This technology requires only 100 ng of RNA to determine quality, with

Table 4 Genes comprising the 30-dimensional data set

A. Genes that appear up-regulated in cancer tissue

Gene	C:NC[a]	P			
		t test[b]	Unequal	*t* test $\log_{10}$	Wilcoxon
BTF-2	3.2	0.010	0.026	0.001	0.008
p160	3.1	0.016	0.030	0.027	0.095
spr2	3.0	0.006	0.013	0.007	0.008
Interferon inducible 9-27	2.9	0.007	0.019	0.003	0.008
Human surface antigen	2.7	0.018	0.035	0.023	0.056
Grb14	2.6	0.004	0.011	0.001	0.008
gp250 precursor	2.6	0.008	0.009	0.002	0.016
TAK1-binding protein	2.4	0.014	0.010	0.002	0.008
Myosin-binding protein H	2.3	0.016	0.022	0.011	0.032
RP3	2.3	0.006	0.008	0.004	0.016
α-Catenin	2.3	0.001	0.004	0.001	0.008
T3 receptor cofactor-1	2.3	0.005	0.006	0.015	0.016
DAP-3	2.2	0.017	0.021	0.025	0.032
Selenoprotein-W (selW)	2.1	0.015	0.023	0.018	0.056
Ferroxidase	2.1	0.007	0.011	0.008	0.016
Cytochrome c	2.1	0.017	0.017	0.044	0.032
RAN-BP8	2.1	0.017	0.018	0.016	0.032
Aspartate aminotransferase-1	1.9	0.010	0.011	0.009	0.032
unc-18 homologue	1.9	0.007	0.010	0.005	0.008
Phosphethanolamine cytidylyltransferase	1.9	0.007	0.007	0.001	0.016
Frezzled (fre)	1.9	0.016	0.024	0.014	0.016
Interferon α-induced 11.5 kDa	1.8	0.017	0.019	0.017	0.032
Ubiquitin activating enzyme E1	1.8	0.015	0.015	0.011	0.032
Macrophage-stimulating 1	1.8	0.014	0.014	0.022	0.032
Ah receptor	1.8	0.002	0.004	0.002	0.008

B. Genes that appear up-regulated in noncancer tissues

Gene	C:NC[a]	P			
		t test[b]	Unequal	*t* test $\log_{10}$	Wilcoxon
Neurofibromin 2	0.6	0.005	0.005	0.006	0.008
Frizzled-related protein	0.5	0.016	0.025	0.015	0.031
Type II keratin	0.4	0.006	0.007	0.027	0.008
CAGH4	0.4	0.004	0.006	0.003	0.008
Dihydroguanosine triphosphatase	0.3	0.014	0.033	0.002	0.008

[a] C:NC, ratio of expression level in cancer *versus* noncancer. Genes were selected on the basis of C:NC $\geq$ 1.8 (approximately 2-fold); $P \leq 0.02$ (estimated to three significant figures) in Student's *t* test.

[b] *t* tests used are: *t* test, Student's (untransformed data); Unequal, unequal variance (untransformed data); *t* test $\log_{10}$, Student's *t* test on $\log_{10}$-transformed data; Wilcoxon, Wilcoxon rank-sum test (nonparametric).

specificity comparable with or better than that obtained from standard gel electrophoresis. Consequently, RNA quality can be assessed on samples that will later be subjected to gene microarray analysis. As is evident from Fig. 1*B*, >90% of representative biopsies produced high-quality RNA.

Analysis of Core Needle Breast Biopsies and Visualization of Gene Expression Data. To assess the applicability of the tissue processing procedure, we obtained total RNA from five random breast cancer biopsies and five random biopsies of noncancer tissue (Table 3). All tissues were evaluated by the study pathologist before release for our studies to ensure that the investigational cores contained no diagnostically useful information. Both biopsies processed in RNA*Later* and biopsies frozen without RNA*Later* were analyzed. These biopsies were approximately equally represented in each group (RNA*Later* processed: cancer = 3; noncancer = 3). RNA was prepared, and probes were generated as described above. The mean RNA concentrations recovered by both methods were comparable (see also Table 2). Probes were hybridized to NamedGene filters, and signal was measured using a Molecular Dynamics Storm PhosphorImager. Digitized representations of the hybridized filter signals were imported into the Pathways software for background correction and normalization.

Normalized gene expression data were imported into the visualization algorithm, and scatter plots of the gene expression data were generated. We first reduced dimensionality by eliminating noninformative genes. Hence, we excluded those genes whose expression was not likely to be different between the cancer and noncancer groups (multiple *t* tests, $P > 0.05$). A total of 103 genes met this criterion and were used to generate a three-dimensional (from 103-dimensional) plot of the data (Fig. 5*A*). The three axes are the first three principal components fitted to the cancer and noncancer molecular profile data. The cumulative proportion of the variance captured by each principal component axis is: (*a*) principal component axis 1, 55%; (*b*) principal component axis 2, 72%; and (*c*) principal component

axis 3, 79%. We also applied hierarchical clustering, similar to approaches used by others (26), based on Euclidean space analysis (1-Pearson's correlation coefficient matrix). The latter approach could not completely separate two cancers from the clusters of noncancers (Fig. 5*B*). PCA[6]-based multidimensional scaling visualization separated breast cancers (*triangles*) and noncancer tissue (*circles*) into linearly separable gene expression data space. However, it should be noted that neither approach provides a statistical assessment of separability, only a visualization of data structure. Whereas the number of data points is limited, the multidimensional scaling visualization is consistent with our ability to identify a putative molecular profile that can separate neoplastic from nonneoplastic tissues.

This subset of genes is expected to include some false positives, reflecting the type 1 error associated with the selection. Consequently, data dimensionality was further reduced using more conservative criteria ($P \leq 0.02$ and regulation $\geq$1.8-fold). We chose this fold regulation to include all $\leq$2-fold differences in mean gene expression levels between cancer and noncancer tissues. The analysis produced a 30-dimensional data set; 25 signals (genes) were up-regulated in the neoplastic biopsies (Table 4*A*), and 5 signals were up-regulated in the nonneoplastic biopsies (Table 4*B*). The ability of this subset to separate cancer from noncancer was also evaluated using both our PCA-based multidimensional scaling approach and simple hierarchical clustering. The cumulative proportion of the variance captured by each principal component axis is: (*a*) principal component axis 1, 64%; (*b*) principal component axis 2, 75%; and (*c*) principal component axis 3, 82%. Neoplastic and nonneoplastic tissues (Table 3) were now linearly separable in gene expression data space by both visualization methods (Fig. 6, *A* and *B*).

Neural Network Predictors of Biopsy Phenotypes. Having reduced the dimensionality, it was necessary to assess whether the expression patterns of remaining genes in the 103 and 30 dimensions contained useful discriminatory information. Thus, the ability of various gene subsets to train accurate neural network predictors that could predominantly separate cancer from noncancer tissues was assessed. The three configurations tested (1–3 hidden nodes) for genes within the 30- and 103-dimensional data sets are described in "Materials and Methods." All were evaluated using the leave-one-out method. Whereas the number of microarrays from which the data are obtained is small ($n = 10$), each configuration achieved a 0% misclassification rate (network training) for cancer *versus* noncancer, whether in 103 or 30 dimensions and with either $\log_{10}$ or nontransformed gene expression data.

Because the initial training and testing were done on the original data set from the Georgetown University samples, we tested the neural networks against an independent data set of nine frozen breast specimens from the University of Edinburgh. These were snap-frozen mastectomy specimens rather than core needle breast biopsies, but they should contain a mixture of cancer and noncancer cells and provide a strong and independent challenge for the neural network. The neural network model should accurately predict as cancer any biopsy comprising >80% cancer tissue.

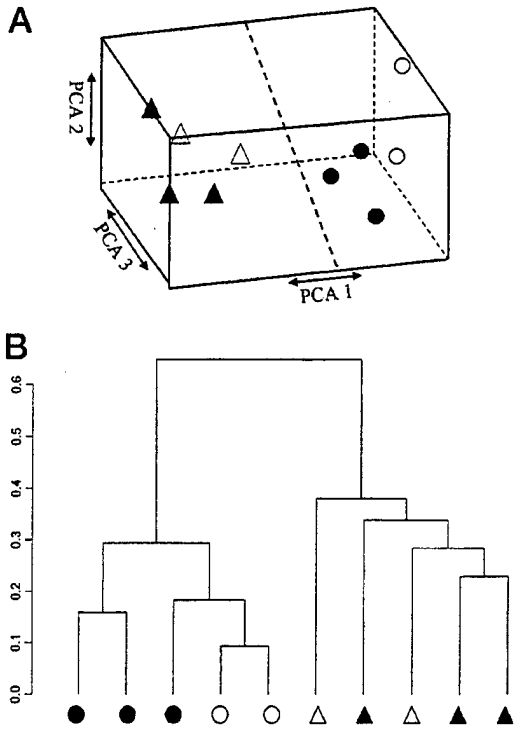


Fig. 6 Structure of the gene expression data using the top 30 genes. *A*, projection from top 30 genes (30 dimensions), selected by *t* test comparisons of $\log_{10}$-transformed gene expression data ($P \leq 0.02$; $\geq$1.8-fold difference) and projected into three dimensions; *B*, hierarchical clustering of samples in 2 dimensions based on 1-Pearson's coefficient matrix of the top 30 genes. Snap-frozen tissue: ○, noncancer; △, cancer; RNA*Later*-processed tissue: ●, noncancer; ▲, cancer.

Gene expression data were generated using the same Research Genetics filter technology and queried in the predictive model. For both the 103 and 30 gene data sets, the nontransformed data provided the more accurate models. Both models predicted that all nine samples should be cancer and not noncancer. The pathologist who evaluated the samples for the training set subsequently performed histopathological analysis of stored samples of these tissues. All nine samples were confirmed as $\geq$80% cancer specimens. Thus, no samples in the independent test data set were misclassified, demonstrating the neural network's predictive accuracy. When the $\log_{10}$ data were used, the models misclassified 1 of 9 tumors (30 dimensions; 89% accurate) and 2 of 9 tumors (103 dimensions; 78% accurate). The lower classification rate with the 103 genes probably reflects the increased type 1 error associated with this data set and the failure to exclude some uninformative genes.

Genes Differentially Expressed between Breast Cancer and Noncancer Tissues. The data in Table 4 show that the choice of *t* test has only a marginal effect on data selection for supervised dimension reduction. If we make no assumption regarding distribution of the data, approximately 1 in 3 genes

[6] The abbreviations used are: PCA, principal component analysis; ER, estrogen receptor.

Table 5 Function of selected genes

Gene name(s)	UniGene no.[a]	Function	Ref. no.
BTF2 (butyrophilin)	Hs.167741	Glycoprotein component of human milk fat globule membranes; membrane-associated receptor for association of cytoplasmic droplets with the apical plasma membrane	35 and 41
grb14 (growth factor receptor-bound protein 14)	Hs.83070	Member of the grb7 family; phosphorylated by a PDGF-regulated serine kinase; expression correlates with ER expression	37
TAB1; TAK1 (MAPKKK; TGFβ-activated kinase-1) *binding protein-1*	Hs.31472	Stimulates NFκB activation; implicated in signaling in response to TGF-β and TNF-α; activates plasminogen activator inhibitor 1	42 and 43
α-*Catenin*	Hs.178452	Cell adhesion molecule; binds E-cadherin; associated with tumor grade and ER expression	38 and 44
DAP3 (death-associated protein-3)	Hs.159627	Proapoptotic, nucleotide-binding protein	45 and 46
Ceruloplasmin; ferroxidase	Hs.28896	Copper transport protein; present in breast milk; serum levels are elevated in breast cancer patients with progressive disease but not in patients in remission or those with benign breast lesions	33, 34, and 47
Ah receptor (aryl hydrocarbon receptor)	Hs.172287	Binds environmental toxins; interacts with ER; can block the transcriptional activity of ER; binds ER co-regulators ERAP140 and SMRT	36, 48, and 49
NF2 (neurofibromatosis-2)	Hs.902	Tumor suppressor; rarely mutated in breast cancer	50 and 51
Frizzled-related gene	Hs.7306	Secreted protein; lost in ~80% of breast cancers; apoptosis related gene, induced by Adriamycin	52 and 53

[a] The UniGene databases can be found at http://www.ncbi.nlm.nih.gov/UniGene/Hs.Home.html.
[b] PDGF, platelet-derived growth factor; NFκB, nuclear factor κB; TGF, transforming growth factor; TNF, tumor necrosis factor.

would be rejected by relying solely on the nonparametric analyses, a ≥1.8-fold differential expression, and a cutoff of $P \leq 0.02$. The 30 target cDNAs comprising the 30-dimensional data set are presented in Table 4.

DISCUSSION

Generally, prospective study designs are more valid for the exploration or validation of new predictive and prognostic factors. Retrospective breast cancer studies may be compromised by the bias toward larger tumors in many existing frozen tumor banks, whereas the average size of most newly diagnosed breast tumors continues to decrease (9). Thus, many studies into the molecular biology of such early lesions may need to be done prospectively. Investigators at single academic institutions can often prospectively obtain frozen samples under a rigorous collection protocol. However, the ability to do so at multiple institutions or when local clinics and community physicians are also involved can be problematic. A rapid, standard tissue processing approach should allow for the use of tissues from multiple institutions in a controlled manner. For example, it should be possible to reduce possible changes in molecular profiles associated with differences in tissue acquisition and processing. Whereas these concerns have not been explored in detail experimentally, tissue processing clearly affects the performance of other molecular biological technologies applied to human biopsies and tumor tissues (10).

To address these issues, we conducted studies to identify an optimal tissue acquisition, processing, and analysis procedure for exploring the gene expression profiles of prospectively accrued breast core needle biopsies. Because RNA extraction destroys tissue architecture, we developed a novel method for tissue processing that would allow us to obtain samples in a uniform manner, preserve RNA quality/quantity, and, most importantly, retain all potentially diagnostically relevant information.

Tissue placed in RNA*Later* can be left at room temperature for up to 1 h at 37°C, 1 week at 25°C, and ≥1 month at 4°C and retain fully intact RNA (31). Our data show that biopsies processed immediately in either liquid nitrogen or RNA*Later* can produce sufficient concentrations of high-quality RNA for nylon filter microarray analysis without RNA amplification. This amount of RNA is also adequate for amplification for use with other gene expression microarray technologies (32). If processed carefully, tissue architecture can be maintained from biopsies collected in RNA*Later*. This is clearly important because some small breast lesions can be completely removed by the biopsy procedure. These core biopsies should not be used for studies if critical diagnostic information could be lost. We estimate that, using the approaches described in this study, approximately 90% of suitable core needle breast biopsies should produce sufficient material for gene expression microarray studies.

Our studies demonstrate that the RNA recovered can be used to generate relevant gene expression microarray information. Relevance is evident from our abilities to identify differentially expressed genes associated with breast cancer cells and to build accurate neural network predictors that can identify cancer from noncancer samples based solely on their gene expression profiles.

Among the differentially expressed genes in the reduced 30-dimensional space, we would expect to find either some genes already implicated in breast cancer or known to be expressed in normal or neoplastic breast tissues. Consistent with this expectation, several genes of potential relevance were iden-

tified. For example, ceruloplasmin is up-regulated in neoplastic breast tissues, and elevated serum levels of ceruloplasmin are associated with recurrent breast cancers (33, 34). The BT2 glycoprotein is a milk protein (35) and might be expected to be expressed in tissues predominately composed of breast epithelial cells.

ER protein status is determined routinely for cancer but not noncancer biopsies. Because four of the five tumor biopsies were ER positive, we also would expect to find genes with expression patterns known either to be associated with ER or to modulate ER function. At least three genes meet these criteria. The aryl hydrocarbon receptor is known to interact with ER and affect its function (36), and the expression of both grb14 and α-catenin is associated with ER expression in breast tumors (Refs. 37 and 38; Table 5).

The discriminant power of the genes selected is evident from the accuracy of the neural networks built using the data from the initial five cancer and five noncancer biopsies. The ability to accurately identify independent samples as cancer shows that the genes of interest are expressed or repressed in both patterns and at levels consistent with the model. This is an appropriate and rigorous test of the approach because the goal was to build molecular predictors, rather than to identify functionally relevant genes. Building a predictor is also a much more efficient test of the selected genes than would be obtained by simply confirming expression gene by gene in more standard assays: Northern blot, RNase protection, or real-time PCR. Confirming the differential expression of each gene is unnecessary for building clinically relevant predictive models. Unlike studies to identify functionally relevant genes, the discriminate power of each signal from the target cDNAs on the array is independent of whether that signal originates from hybridization to its expected mRNA.

The gene expression profile data and neural network performance suggest that, at least for samples of very different biologies, contamination of samples with ≥80% of other cell types may not confound analyses for molecular profiling. Whether this observation can be extrapolated to other studies remains to be further established. Nonetheless, the resource intensive requirements of microdissection and RNA amplification may not be absolute requirements for all molecular profiling studies.

The tissue acquisition and processing methods, dimension reduction, data visualization approaches, and neural network analyses we describe may be useful in the design of larger prospective studies. We continue to develop other data visualization, normalization, and exploration algorithms that also may be of use in the analysis of gene expression microarray studies (24, 25, 39, 40).

REFERENCES

1. Hayes, D. F., Bast, R. C., Desch, C. E., Fritsche, H., Kemeny, N. E., Jessup, J. M., Locker, G. Y., MacDonald, J. S., Mennel, R. G., Norton, L., Ravdin, P. M., Taube, S., and Winn, R. J. Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. J. Natl. Cancer Inst. (Bethesda), *88:* 1456–1466, 1996.

2. Hayes, D. F., Trock, B., and Harris, A. Assessing the clinical impact of prognostic factors: when is "statistically significant" clinically useful? Breast Cancer Res. Treat., *52:* 305–319, 1998.

3. Marx, J. DNA arrays reveal cancer in its many forms. Science (Wash. DC), *289:* 1670–1672, 2000.

4. Carulli, J. P., Artinger, M., Swain, P. M., Root, C. D., Chee, L., Tulig, C., Guerin, J., Osborne, M., Stein, G., Lian, J., and Lomedico, P. T. High throughput analysis of differential gene expression. J. Cell. Biochem., *30–31:* 286–296, 1998.

5. Richmond, C. S., Glasner, J. D., Mau, R., Jin, H., and Blattner, F. R. Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res., *27:* 3821–3835, 1999.

6. Cox, J. M. Applications of nylon membrane arrays to gene expression analysis. J. Immunol. Methods, *250:* 3–13, 2001.

7. Bertucci, F., Bernard, K., Loriod, B., Chang, Y. C., Granjeaud, S., Birnbaum, D., Nguyen, C., Peck, K., and Jordan, B. R. Sensitivity issues in DNA array-based expression measurements and performance of nylon microarrays for small samples. Hum. Mol. Genet., *8:* 1715–1722, 1999.

8. Granjeaud, S., Bertucci, F., and Jordan, B. R. Expression profiling: DNA arrays in many guises. Bioessays, *21:* 781–790, 1999.

9. Morrow, M., Schnitt, S. J., and Harris, J. R. *In situ* carcinomas. *In:* J. R. Harris, M. E. Lippman, M. Morrow, and S. Hellman (eds.), Diseases of the Breast, pp. 355–373. Philadelphia: Lippincott-Raven, 1996.

10. Zeller, R. *In situ* hybridization and immunohistochemistry. *In:* F. M. Ausbel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl (eds.). Current Protocols in Molecular Biology, pp. 14.1.1–14.14.8. New York: John Wiley & Sons, Inc., 2001.

11. Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. Normalization strategies for cDNA microarrays. Nucleic Acids Res., *28:* E47, 2000.

12. Wang, G. L., and Semenza, G. L. General involvement of hypoxia-inducible factor 1 in transcriptional response to hypoxia. Proc. Natl. Acad. Sci. USA, *90:* 4304–4308, 1993.

13. Wenger, R. H., Rolfs, A., Marti, H. H., Bauer, C., and Gassmann, M. Hypoxia, a novel inducer of acute phase gene expression in a human hepatoma cell line. J. Biol. Chem., *270:* 27865–27870, 1995.

14. Leonessa, F., Green, D., Licht, T., Wright, A., Wingate-Legette, K., Lippman, J., Gottesman, M. M., and Clarke, R. MDA435/LCC6 and MDA435/LCC6^{MDR1}: ascites models of human breast cancer. Br. J. Cancer, *73:* 154–161, 1996.

15. Clarke, R. Human breast cancer cell line xenografts as models of breast cancer: the immunobiologies of recipient mice and the characteristics of several tumorigenic cell lines. Breast Cancer Res. Treat., *39:* 69–86, 1996.

16. Sgroi, D. C., Teng, S., Robinson, G., LeVangie, R., Hudson, J. R., and Elkahloun, A. G. *In vivo* gene expression profile analysis of human breast cancer progression. Cancer Res., *59:* 5656–5661, 1999.

17. Walker, J., and Rigley, K. Gene expression profiling in human peripheral blood mononuclear cells using high-density filter-based cDNA microarrays. J. Immunol. Methods, *239:* 167–179, 2000.

18. McCormick, S. M., Eskin, S. G., McIntire, L. V., Teng, C. L., Lu, C. M., Russell, C. G., and Chittur, K. K. DNA microarray reveals changes in gene expression of shear stressed human umbilical vein endothelial cells. Proc. Natl. Acad. Sci. USA, *98:* 8955–8960, 2001.

19. Herwig, R., Aanstad, P., Clark, M., and Lehrach, H. Statistical evaluation of differential expression on cDNA nylon arrays with replicated experiments. Nucleic Acids Res., *29:* E117, 2001.

20. Carlisle, A. J., Prabhu, V. V., Elkahloun, A., Hudson, J., Trent, J. M., Linehan, W. M., Williams, E. D., Emmert-Buck, M. R., Liotta, L. A., Munson, P. J., and Krizman, D. B. Development of a prostate cDNA microarray and statistical gene expression analysis package. Mol. Carcinog., *28:* 12–22, 2000.

21. Hinneburg, A., and Keim, D. A. Optimal grid-clustering: toward breaking the curse of dimensionality in high-dimensional clustering. *In:* M. P. Atkinson, M. E. Orlowska, P. Valduriez, S. B. Zdonik, and M. L. Brodie (eds.), Proceedings of the 25th Conference on Very Large Databases, pp. 506–517. San Francisco: Morgan Kaufman, 1999.

22. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., and Trent, J. Gene-expression profiles in hereditary breast cancer. N. Engl. J. Med., *344:* 539–548, 2001.

23. Wittes, J., and Friedman, H. P. Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data. J. Natl. Cancer Inst. (Bethesda), *91:* 400–401, 1999.

24. Wang, Y., Lin, S. H., Li, H., and Kung, S. Y. Data mapping by probabilistic modular networks and information theoretic criteria. IEEE Trans. Signal Processing, *46:* 3378–3397, 1998.

25. Wang, Y., Luo, L., Freedman, M. T., and Kung, S. Y. Probabilistic principal component subspaces: a hierarchical finite mixture model for data visualization. IEEE Trans. Neural Net., *11:* 635–646, 2000.

26. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, *95:* 14863–14868, 1998.

27. Haykin, S. Neural Networks: A Comprehensive Foundation. Upper Saddle River, NJ: Prentice Hall, Inc., 1999.

28. Jain, A. K., Duin, R. P. W., and Mao, J. Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Machine Intel., *22:* 4–37, 2000.

29. Adali, T., Wang, Y., and Li, H. Neural networks for biomedical signal processing. *In:* Y-H. Hu and J-N. Huang (eds.), Handbook of Neural Network Signal Processing, in press. Boca Raton, FL: CRC Press, Inc., 2001.

30. Ripley, B. Pattern Recognition and Neural Networks. Cambridge, U.K.: Cambridge University Press, 1996.

31. RNALater Tissue Collection. RNA Stabilization Solution. *In:* Ambion Protocols and Manuals. Austin, TX: Ambion, 2001.

32. Eberwine, J. Amplification of mRNA populations using aRNA generated from immobilized oligo(dT)-T7 primed cDNA. Biotechniques, *20:* 584–591, 1996.

33. Ozyilkan, O., Baltali, E., Ozyilkan, E., Tekuzman, G., Kars, A., and Firat, D. Ceruloplasmin level in women with breast disease. Preliminary results. Acta Oncol., *31:* 843–846, 1992.

34. Schapira, D. V., and Schapira, M. Use of ceruloplasmin levels to monitor response to therapy and predict recurrence of breast cancer. Breast Cancer Res. Treat., *3:* 221–224, 1983.

35. Peterson, J. A., Hamosh, M., Scallan, C. D., Ceriani, R. L., Henderson, T. R., Mehta, N. R., Armand, M., and Hamosh, P. Milk fat globule glycoproteins in human milk and in gastric aspirates of mother's milk-fed preterm infants. Pediatr. Res., *44:* 499–506, 1998.

36. Klinge, C. M., Kaur, K., and Swanson, H. I. The aryl hydrocarbon receptor interacts with estrogen receptor α and orphan receptors COUP-TFI and ERRα1. Arch. Biochem. Biophys., *373:* 163–174, 2000.

37. Daly, R. J., Sanderson, G. M., Janes, P. W., and Sutherland, R. L. Cloning and characterization of GRB14, a novel member of the *GRB7* gene family. J. Biol. Chem., *271:* 12502–12510, 1996.

38. Gonzalez, M. A., Pinder, S. E., Wencyk, P. M., Bell, J. A., Elston, C. W., Nicholson, R. I., Robertson, J. F., Blamey, R. W., and Ellis, I. O. An immunohistochemical examination of the expression of E-cadherin, α- and β/γ-catenins, and α_2- and β_1-integrins in invasive breast cancer. J. Pathol., *187:* 523–529, 1999.

39. Wang, Y., Lu, J., Lee, R. Y., and Clarke, R. Iterative normalization of cDNA microarray data. IEEE Trans. Inf. Technol. Biomed., *6:* 29–37, 2002.

40. Lu, J., Wang, Y., Xuan, J., Kung, S. Y., Gu, Z., and Clarke, R. Discriminative mining of gene microarray data. Proc. IEEE Neural Net Signal Processing, *11:* 218–227, 2001.

41. Heid, H. W., Winter, S., Bruder, G., Keenan, T. W., and Jarasch, E. D. Butyrophilin, an apical plasma membrane-associated glycoprotein characteristic of lactating mammary glands of diverse species. Biochim. Biophys. Acta, *728:* 228–238, 1983.

42. Sakurai, H., Miyoshi, H., Toriumi, W., and Sugita, T. Functional interactions of transforming growth factor β-activated kinase 1 with IκB kinases to stimulate NF-κB activation. J. Biol. Chem., *274:* 10641–10648, 1999.

43. Shibuya, H., Yamaguchi, K., Shirakabe, K., Tonegawa, A., Gotoh, Y., Ueno, N., Irie, K., Nishida, E., and Matsumoto, K. TAB1: an activator of the TAK1 MAPKKK in TGF-β signal transduction. Science (Wash. DC), *272:* 1179–1182, 1996.

44. Bukholm, I. K., Nesland, J. M., and Borresen-Dale, A. L. Reexpression of E-cadherin, α-catenin and β-catenin, but not of γ-catenin, in metastatic tissue from breast cancer patients. J. Pathol., *190:* 15–19, 2000.

45. Berger, T., Brigl, M., Herrmann, J. M., Vielhauer, V., Luckow, B., Schlondorff, D., and Kretzler, M. The apoptosis mediator mDAP-3 is a novel member of a conserved family of mitochondrial proteins. J. Cell Sci., *113:* 3603–3612, 2000.

46. Levy-Strumpf, N., and Kimchi, A. Death-associated proteins (DAPs): from gene identification to the analysis of their apoptotic and tumor suppressive functions. Oncogene, *17:* 3331–3340, 1998.

47. Kunapuli, S. P., Singh, H., Singh, P., and Kumar, A. Ceruloplasmin gene expression in human cancer cells. Life Sci., *40:* 2225–2228, 1987.

48. Trombino, A. F., Near, R. I., Matulka, R. A., Yang, S., Hafer, L. J., Toselli, P. A., Kim, D. W., Rogers, A. E., Sonenshein, G. E., and Sherr, D. H. Expression of the aryl hydrocarbon receptor/transcription factor (AhR) and AhR-regulated *CYP1* gene transcripts in a rat model of mammary tumorigenesis. Breast Cancer Res. Treat., *63:* 117–131, 2000.

49. Nguyen, T. A., Hoivik, D., Lee, J. E., and Safe, S. Interactions of nuclear receptor coactivator/corepressor proteins with the aryl hydrocarbon receptor complex. Arch. Biochem. Biophys., *367:* 250–257, 1999.

50. Yaegashi, S., Sachse, R., Ohuchi, N., Mori, S., and Sekiya, T. Low incidence of a nucleotide sequence alteration of the *neurofibromatosis 2* gene in human breast cancers. Jpn. J. Cancer Res., *86:* 929–933, 1995.

51. Kanai, Y., Tsuda, H., Oda, T., Sakamoto, M., and Hirohashi, S. Analysis of the neurofibromatosis 2 gene in human breast and hepatocellular carcinomas. Jpn. J. Clin. Oncol., *25:* 1–4, 1995.

52. Preiherr, J., Hildebrandt, T., Klostermann, S., Eberhardt, S., Kaul, S., and Weidle, U. H. Transcriptional profiling of human mammary carcinoma cell lines reveals PKW, a new tumor-specific gene. Anticancer Res., *20:* 2255–2264, 2000.

53. Ugolini, F., Adelaide, J., Charafe-Jauffret, E., Nguyen, C., Jacquemier, J., Jordan, B., Birnbaum, D., and Pebusque, M. J. Differential expression assay of chromosome arm 8p genes identifies Frizzled-related (FRP1/FRZB) and fibroblast growth factor receptor 1 (FGFR1) as candidate breast cancer genes. Oncogene, *18:* 1903–1910, 1999.